Universiteit Gent
Faculteit Wetenschappen
Statistical Data Analysis

Low-Frequency Variant Detection in Viral Populations using Massively Parallel Sequencing Data

Detectie van laag frequente varianten in virale populaties gebruik makend van massaal parallel sequentioneringsdata

---

Bie Verbist

Proefschrift tot het bekomen van de graad van
Doctor in de Statistische Data Analyse
Academiejaar 2014-2015

Universiteit Gent
Faculteit Wetenschappen
Statistical Data Analysis

Academische Promotoren: Prof. Dr. Ir. Olivier Thas
Prof. Dr. Ir. Lieven Clement
Industriële Promotor: Prof. Dr. Luc Bijnens

Universiteit Gent
Faculteit Wetenschappen

Statistical Data Analysis
Krijgslaan 281, S9,
B-9000 Gent, België

Tel.: +32-9-264.47.57
Fax.: +32-9-264.49.95

Dit werk kwam tot stand in het kader van een baekelandmandaat (BM100679) van het IWT-Vlaanderen (Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen).

Proefschrift tot het bekomen van de graad van
Doctor in de Statistische Data Analyse
Academiejaar 2014-2015

Dankwoord

Negen jaar geleden, had ik nooit vermoed dat ik opnieuw een proefschrift zou indienen. Sinds mijn doctoraat in de scheikunde heeft mijn bescheiden carrière al een hele ommezwaai gemaakt, hoewel er steeds een constante gebleven is: "puzzelen met data". Tijdens de scheikunde opleiding en mijn post-doc was het meer het opstellen van hypotheses, gebaseerd op biologische data voor een optimaal design van de molecules. Ondertussen hebben de begrippen hypothese en design een meer statistische invulling gekregen.

Ook deze keer zou het niet gelukt zijn, zonder de steun van verschillende mensen. In de eerste plaats wil ik Els Goetghebeur bedanken omdat zij mij aangemoedigd heeft om contact op te nemen met Janssen Pharmaceutica in mijn zoektocht naar een thesis onderwerp voor de master of statistical data analysis. Zo ben ik uiteindelijk in contact gekomen met Luc Bijnens. Hij heeft mij niet alleen een thesis onderwerp aangeboden, maar heeft ook de basis gelegd voor het huidige project. Luc, bedankt om mij deze kansen te bieden. Jij geeft mij altijd die extra stimulans waardoor ik continu mijn eigen grenzen weet te verleggen. Aan de academische zijde wens ik mijn promotoren Olivier Thas en Lieven Clement te bedanken. Olivier, bedankt voor al uw inzet om de verschillende research vragen te vertalen naar statistische concepten en deze dan ook nog op een begrijpbare manier uit te leggen. Lieven, soms was het vloeken als je weer met nieuwe ideëen aan kwam draven, maar de kwaliteit van het werk verbeterde aanzienlijk. Bedankt daarvoor. Jullie beiden hebben mij, als wetenschapper, verder ontplooid om ook in statistische termen te denken. Verder wens ik ook Jeroen, Kim en Joke te bedanken. Tijdens onze tweemaandelijkse ‘sequencing statistics’ meetings waren jullie steeds kritisch maar constructief zodat de toepasbaarheid van de methodes verzekerd bleef. Het was zeer verrijkend om in deze multidisciplinaire teams te mogen meedraaien. Yves, uren hebben wij samen aan de computer doorgebracht om alles geprogrammeerd te krijgen en nooit was het je teveel om nog maar eens een kleine aanpassing door te voeren. Bedankt voor deze constante inzet. Tobias, Joris en Alexander thanks to give the necessary IT and programming support. Verder wil ik ook alle mensen binnen discovery sciences die bewust of onbewust tot het project hebben bijgedragen bedanken. In het bijzonder, wil ik de mensen van het computational biology team binnen deze groep een woord van dank toewerpen. Het was altijd leuk om op dinsdag bij jullie te mogen vertoeven.

Het is duidelijk dat dit project gedragen werd zowel langs academische als industriële zijde. Toch is het niet altijd gemakkelijk om tussen deze twee werelden te opereren. Daarom wil ik de twee teams bedanken waar ik deel van uitmaakte. In de eerste plaats, het niet-klinische statistics team in Beerse. Meestal zat ik fysiek bij jullie. Bedankt om mij, weliswaar als aanhangsel ;-), in jullie team op te nemen. Willem wil ik in het bijzonder bedanken omdat hij me bij de start van het project wegwijs gemaakt heeft binnen het bedrijf. Aan de andere kant wil ik alle collega's van BioStat bedanken. Die keren dat ik in Gent was, stonden jullie altijd wel open voor een fijne babbel. Speciaal woord van dank aan Federico. Thanks to keep me updated about what was happening at the university whenever you were visiting Janssen. Bedankt ook aan Timpe, Annie, Roos, Ruth en Katrien voor alle administratieve support.

Natuurlijk mag ik in dit lijstje ook niet vergeten de financiële sponsors te bedanken. Dit project werd goedgekeurd als een Baekeland project door het IWT en werd gecofinancierd door Janssen Pharmaceutica. Luc bedankt om binnen Janssen hier budget voor te voorzien.

Verder wil ik mijn familie en vrienden bedanken. Zij zorgen er voor dat ik telkens weer de nieuwe energie kan vinden om verder te gaan. Speciaal woord van dank aan mijn ouders. Zij zijn de meest trouwe supporters en staan altijd klaar om in te springen als ik nog maar eens handen te kort kom. Het allerlaatste woord van dank hou ik voor de twee mannen in mijn leven: Raf en Tuur. Raf, altijd sta je klaar om me op te beuren, maar evengoed om mee kleine overwinningen te vieren. Bedankt daarvoor. Tuur, jij met je onuitputtelijke energie slaagt er steeds in om mij al de zorgen eventjes te doen vergeten. Tuur bedankt ook om mijn voorpagina te ontwerpen en ik zal het onthouden, de virussen met de stekels zijn de gevaarlijkste.

Voor dit alles en nog veel meer, bedankt allemaal.

Gent, december 2014
Bie Verbist

Table of Contents

List of Acronyms

A

A	Adenine
AIDS	Acquired Immune Deficiency Syndrome
APS	Adenosine 5′ PhosphoSulfate
ART	Antiretroviral Therapy
ASCII	American Standard Code Information Interchange
ATP	Adenosine TriPhosphate

B

bp	base pair
BWA	Burrows-Wheeler Aligner

C

C	Cytosine
cDNA	Complementary DNA
CLC Bio	Bioinformatics software company; while the acronym behind CLC remains a secret with the founding brothers, it has been suggested that it stands for Cake-Loving Company.

D

DNA	DeoxyriboNucleic Acid
dNTP	DeoxyribonNucleotide TriPhosphate

E

EM	Expectation Maximization

F

FDR	False Discovery Rate

G

G	Guanine
GA	Genome Analyzer
GoF	Goodness of Fit

H

HCV	Hepatitis C Virus
HIV	Human Immunodeficiency Virus

I

IWT	Innovatie door Wetenschap en Technologie

M

MPS	Massive Parallel Sequencing
mRNA	Messenger RNA

N

NGS	Next Generation Sequencing
NS3	Non-Structural protein 3

P

PCR	Polymerase Chain Reaction
PR	Protease

Q

QIT	Quality Intersection Threshold

R

RNA	RiboNucleic Acid
RT	Reverse Transcriptase

S

SMRT	Single Molecule Real Time sequencing
SNP	Single Nucleotide Polymorphism

T

T	Thymine
TM	Trade Mark

Nederlandse samenvatting –Summary in Dutch–

Massaal parallel sequentioneringstechnieken zijn veelbelovend voor het in kaart brengen van virale populaties in bijvoorbeeld HIV-1 en HCV geïnfecteerde patienten. De analyse van deze virale populaties kan ons inzicht geven in de ontwikkeling van resistentie hetgeen in een volgend stadium behandeling kan verbeteren. Standaard genotypering levert enkel informatie over de meest voorkomende virale varianten in de populatie. Massaal parallel sequentioneringstechnieken daarentegen laten toe om ook de laag frequente varianten te typeren. De pyrosequencing techniek, gecommercialliseerd door Roche werd tot voor kort het meest gebruikt voor de detectie van deze laag frequente varianten. Maar de MPS-techniek ontwikkeld door Illumina is aan een opmars bezig en heeft het grote voordeel een grotere sequentioneringsdiepte te bereiken voor dezelfde kostprijs. Bovendien heeft Roche recent aangekondigd dat ze de pyrosequencing techniek niet verdere ondersteunen. Daarom bestudeerden we in eerste instantie of Illumina's techniek inderdaad geschikt is voor de karakterisatie van genetische variabiliteit binnen virale populaties en of de 1% rapporteringstechniek, die gebruikelijk is voor 454, ook toegepast kan worden. We kunnen concluderen dat varianten aanwezig in de populatie met een frequentie tot 1% accuraat gedetecteerd kunnen worden.

Eén van de grootste uitdagingen in de detectie van laag frequente varianten zijn de fouten geïncorporeerd tijdens het sequentioneringsproces. Het onderscheiden van fouten van laag frequente varianten wordt bemoeilijkt doordat beiden kunnen voorkomen aan vergelijkbare frequenties. Vermits deze fouten technologisch niet vermeden kunnen worden, moeten we op zoek gaan naar statistische algoritmes die helpen bij het differentiëren. Idealiter willen we varianten detecteren met frequenties ver onder de 1%. Dit zou ons toelaten om de klinische relevantie van de laag frequente varianten te bestuderen in de context van verschillende anti-retrovirale behandelingsregimes. Twee verschillende wegen werden bewandeld om dit doel te bewerkstelligen.

In eerste instantie werden de kwaliteitsscores, meegeleverd door Illumina tijdens het sequentioneringsproces, gebruikt als filteringscriterium om het aantal vals positieven terug te dringen. De grenswaarde voor de kwaliteitsscores, waaronder we de variant beschouwen als fout, wordt bepaald aan de hand van getrunceerde normale mixture distributies gefit op de kwaliteitsscores. Toepassing van deze methode op zowel klinische stalen als plasmides leert ons dat we het aantal vals positieven inderdaad kunnen terugsschroeven vooral in GC-rijke gebieden, waar

Illumina meer dan een gemiddeld aantal fouten maakt. De 1% rapporteringslimiet kon echter nauwelijks naar beneden gebracht worden.

Uit de literatuur en eigen experimenten leerden we dat de kwaliteitsscores niet altijd de echte error probabiliteit weergeven. Vaak zijn deze een onderschatting. Daarom modelleren we de error probabiliteiten van de beste en tweede beste base calls in functie van de kwaliteitsscores. Deze probabiliteiten geven aan of een bepaalde read afkomstig is van een gegeven cluster. De virale populatie kan bijgevolg afgeleid worden van de cluster centers en de grootte van de cluster. Deze methode laat toe om laag frequente varianten te detecteren waarbij we ver onder de 1% duiken zonder aan specificiteit in te boeten.

Dit doctoraatsprogramma getiteld "detectie van laag frequente varianten in virale populaties gebruik makend van massaal parallel sequentioneringsdata"werd positief beoordeeld door het IWT (Baekeland mandaat 100679) en ging van start op 1 januari 2011. Het project is uitgevoerd in nauwe samenwerking tussen Janssen Pharmaceutica en Universiteit Gent.

English summary

In a virology research environment, massive parallel sequencing technology has great opportunity to study viral quasispecies in HIV-1 and HCV-infected patients, which is essential for understanding pathways to resistance and can substantially improve treatment. Whereas standard genotyping only provides information on the most abundant sequence variants, the massively parallel sequencing technologies allow in-depth characterization of sequence variation in more complex populations, including low-frequency viral strains. Until recently, pyrosequencing platforms, commercialized by Roche, have been the most popular for detection of low-level drug resistant variants. However, current short-read sequencing technologies, like the Illumina's sequencing-by-synthesis platform, have the advantage of providing a higher sequencing depth at a lower cost per sequenced base. Additionally, Roche recently announced that they will fade out the pyrosequencing technique by mid-2016. Hence, we investigated the feasibility of using Illumina's GAIIx to characterize genetic variability in viral populations where the key question was to achieve the same widely accepted lower limit of detection of 1% as in 454. We conclude that variants down to a frequency of 1% could be detected with a great accuracy using Illumina's sequencing platform at a lower cost.

One of the challenges in the detection of low-frequency viral strains concerns the errors introduced during the sequencing process. Technology-associated errors may occur up to equal or even higher frequencies than the truly present mutations, impeding a powerful assessment of low-frequency virus mutations. As there are no obvious solutions to reduce the technical noise by further improvements of the technology, we believe that the search for statistical algorithms that can better correct the technical noise can be pivotal. This has the potential to enable sequencing at a much deeper level, far below the 1% level. Having the desired algorithm at hand we could investigate the relevance of minor mutations in the context of different antiretroviral therapy regimens because they might help in defining the clinical benefit of low-frequency resistance testing. Two different approaches to differentiate technical noise from low-frequency variants were investigated.

At first, we used the sequencing quality scores of Illumina as filtering criteria to reduce the number of false findings during variant calling. These quality scores reflect the probability of an error during sequencing. Instead of applying hard thresholds, which are often too stringent or too relaxed, we developed an adaptive thresholding method based on fitting truncated mixture distributions on the quality scores. With this approach we could reduce the number of false-positive findings, especially in GC-rich regions which are known to be error prone. However, the

1% limit of detection could only partially be lowered.

From literature and own experiments, we know that the quality scores do not always reflect the true error probabilities. Often they are underestimated. Therefore in our second approach we modeled the error probabilities of the best and second best base calls as a function of the quality scores. These probabilities express if a given read was generated by a given cluster. The viral population can be inferred from the cluster centers and the cluster sizes. This approach reduces the number of false-positive findings drastically and allows us to lower the 1% limit of detection without losing the specificity.

This doctoral research program "low-frequency variant detection in viral populations using massively parallel sequencing data" was granted by IWT (Baekeland mandatory 100679) and was performed in close collaboration between Janssen Pharamaceuticals and Ghent University.

1 Introduction

Virology is the study of viruses, infectious agents that reproduce inside the cells of living hosts. Viruses consists of genetic material made from either DNA or RNA which are surrounded by a protective coat. They depend on host cells that they infect to reproduce and they can infect all types of life forms, from animals and plants to bacteria. Since Dmitri Ivanovsky's 1892 article [1] describing a non-bacterial pathogen infecting tobacco plants, viruses are found to be the most abundant biological entities on the planet, most of which infect microorganisms [2]. Recent studies estimated that there is a minimum of 320,000 viruses in mammals [3]. One motivation to study these viruses is the fact that they can cause many infectious diseases. Most viral infections, however, are short-lived and of little consequence to the host. Among the reasons for this is the development of a defense system in the host to inhibit viral replication and to destroy virally infected cells in the body [4]. As these systems have developed in their hosts, viruses have needed to modify themselves in order to evade or subvert this immune response. The host immune response is a very complex interwoven series of chemical and cellular interactions that combine to attempt to eliminate viruses from the body. While it is not surprising that viruses have had to develop strategies to overcome host defenses, the number of adaptations and the complexities of these adaptations is remarkable. Adaptations of some viruses are so successful that they are able to escape from these immune responses and can cause severe chronic infections, like for instance Human immunodeficiency virus (HIV) and hepatitis C (HCV) [5, 6]. HCV infects cells in the liver called hepatocytes which triggers the human immune system and leads to inflammation. However, due to the chronic infection,

these prolonged inflammations cause scarring and extensive scarring in the liver is called cirrhosis. When the liver becomes cirrhotic, it fails to perform its normal functions and this leads to serious complications and even death. HIV on the other hand is able to weaken the immune system itself by destroying important cells, called T-cells or CD4-cells that fight disease and infection. Over time, HIV can destroy so many of these cells that your body can't fight infections and diseases anymore. When that happens, HIV infection can lead to AIDS, the final stage of HIV infection. Patients having AIDS are getting infections or cancers that rarely occur in healthy people, because of the damage of their immune system and these infections can be deadly. It is clear that both diseases represent a significant global health threat [7]. HIV and HCV affect millions of humans worldwide, with estimates of 35 million people living with HIV at the end of 2013 and 150 million people with chronic hepatitis C infection. Each year 1.5 million people die from HIV-related causes globally, while 350,000 to 500,000 people die each year from HCV-related liver diseases.

1.1 Viral dynamics

During the viral life cycle, a virus enters the host cell by binding to its receptor, uncoats, make replicates of its own DNA or RNA and proteins using polymerases (see text box for more background information), and then reassembles to form new virus particles which are subsequently released into the host system (Figure 1.1). When viruses infect host cells they provide some of their own molecular equipment such as DNA/RNA polymerases and proteases to enable replication. However, this still requires that, as soon as the cell is invaded, the virus hijack the host cell machinery to manipulate cellular proteins in order to do their replication. Host cells are forced in this way to produce many thousands of copies of the original virus at an extraordinary rate. For HIV and HCV, the number of copies can reach 10^{11} to 10^{12} per infected individual per day. However, these replications are error-prone, resulting in high mutation rates. Especially RNA-viruses, like HIV and HCV with RNA as genetic material have high mutation rates which is attributed partly to the absence of a proof-reading repair activity in the RNA polymerases in contrast to the DNA polymerases [8, 9] during their replication. These high mutation rates together with short generation times result in a constant production of genetic variants. Consequently, RNA viruses exist in their host as complex populations composed of several closely related subgroups, which are referred to as viral quasi-species (Figure 1.2) [10–12]. This heterogeneous mixture of genomes allows a viral population to rapidly adapt to changing environments; for example after infecting a new host with a different immune response [13] or while being exposed to different drugs [14–17].

Antiviral drugs are developed to target specific parts within the viral life cycle.

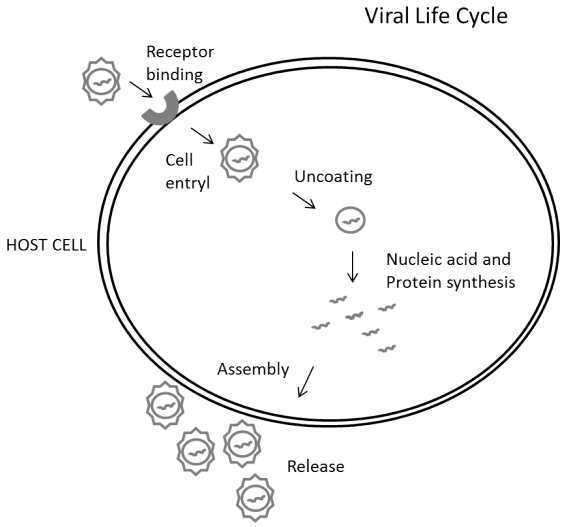


Figure 1.1: General view on the viral life cycle. During the viral life cycle, viruses bind to their receptor, enter the host cell, uncoats, build copies of their genetic material and assemble. At the end copies of virus are released to the host system.

Because viruses use the host's cells to replicate, it is very difficult to find targets for the drug that would interfere with the virus without also harming the host organism's cells. Hence, often the targets are proteins that are translated from the viral genome. However, viral variation imposes a clear challenge here. Viruses carrying mutations in parts of the viral genome which translate to the drug-targets might develop resistance to the antiviral drugs and will out compete other viral variants. This is of course an important threat in the treatment of patients where drug resistance variants are present within the viral population. According to epidemiological studies in Europe and the USA, 8% to 11% of antiretroviral naive patients, patients that have never been treated with relevant drugs against HIV, are infected with a virus harboring drug resistance variants [18]. Therefor, therapeutic guidelines suggest that treatment management at baseline can be improved by a more detailed characterization of sequence variation within the viral population present in a patient. The standard of care for HIV patients for instance is a combination therapy with 2 to 3 different classes of antiretroviral drugs, attacking the viruses in different stages of their viral life cycle. If a person's strain of HIV is resistant to a certain class of drug, taking that type of drug may be ineffective or at worse, harmful as it may lead to failure of the treatment. Hence this type of drug should be avoided in the combination therapy. Not only at the start of the treatment, but also during the treatment, sequence variation should be investigated. As a result of poor adherence, interruptions in treatment and the use of ineffective

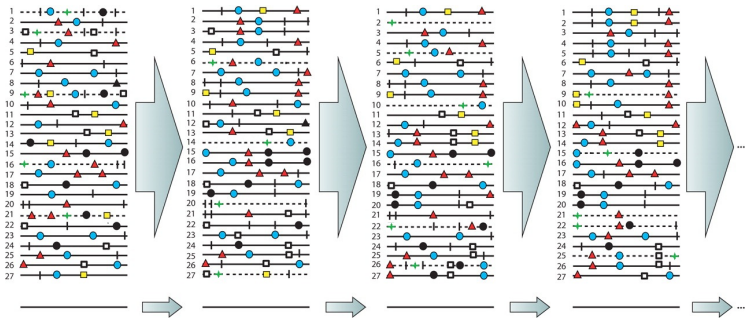


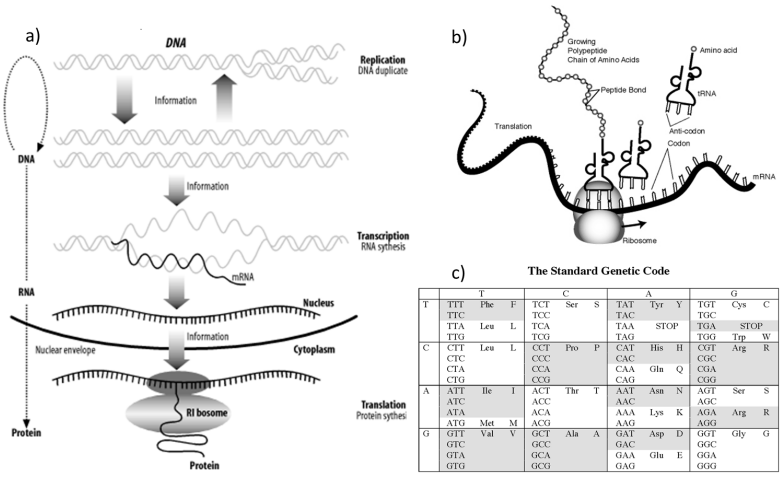
Figure 1.2: Evolution of viral quasispecies. During replication mutations are incorporated in the genomes, resulting in a viral population consisting of closely related subgroups. Viral genomes are represented as horizontal lines and mutations as different colored symbols on the lines. Discontinuous lines indicate genomes that have acquired deleterious mutations, indicated as a green star. These variants cannot survive. (adapted from http://mmbr.asm.org/content/76/2/159/F5.expansion.html)

drugs or faulty drug combinations, HIV can acquire resistance associated mutations during treatment. When the resistance develops, usually the drug regimen needs to be changed.

The pre-existence of these variants, prior to treatment, can be caused as said by the continuous input of new genomic sequences during the replication or is transmitted during infection. It can be predicted that all possible mutations naturally occur at a frequency of $\approx 0.01\%$ [19] since the polymerases, involved in the replication, make one error in 10^4 to 10^5 nucleotides. Therefore, only those variants that occur at higher frequencies are likely to be able to out compete the other variants to develop resistance. A recent study suggest that low-frequency resistant variants in HIV [20] are only clinically meaningful when present above 1% in the viral population. However, it remains a debated issue whether the presence of minority sub-populations that decrease the sensitivity to anti(retro)viral agents influence the treatment outcome or not and which low-frequencies are still clinically relevant. According to some studies, low-frequency baseline drug resistance is associated with a higher risk of treatment failure [14–16]. Other studies have not found an influence of minority resistance mutations on the treatment response [21].

Currently, in clinical practice the determination of the variants within the viral population, called genotyping is done by Sanger sequencing [22]. Sanger sequencing can only detect viral variants representing more than 15 to 25% of the viral population. Although, genotyping can be performed through a variety of different methods, sequencing has the added value that no prior knowledge of the variants is needed. Massively parallel sequencing (MPS) technologies allow a in-depth characterization of viral populations, including low-frequency viral strains [23].

Central dogma of molecular biology.



	T			C			A			G		
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C
	TTC			TCC			TAC			TGC		
	TTA	Leu	L	TCA			TAA	STOP		TGA	STOP	
	TTG			TCG			TAG			TGG	Trp	W
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R
	CTC			CCC			CAC			CGC		
	CTA			CCA			CAA	Gln	Q	CGA		
	CTG			CCG			CAG			CGG		
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S
	ATC			ACC			AAC			AGC		
	ATA			ACA			AAA	Lys	K	AGA	Arg	R
	ATG	Met	M	ACG			AAG			AGG		
G	GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G
	GTC			GCC			GAC			GGC		
	GTA			GCA			GAA	Glu	E	GGA		
	GTG			GCG			GAG			GGG		

Figure 1.3: a) Flow of the genetic information within a biological system from DNA replication, transcription of DNA to RNA and translation of RNA to proteins. b) Translation of genetic information to protein level is done by nucleotide triplets which code for the different amino acids. c) The genetic code of the 64 (4^3) possible codons and their corresponding amino acids. (source http://en.wikipedia.org/wiki/File:Translation-genetics.png)

The central dogma is an explanation of the flow of genetic information within a biological system [24] (Figure 1.3a). The general transfers consist of (1) DNA replication, (2) transcription of DNA to RNA, and (3) translation of RNA to proteins. (1) Deoxyribonucleic acid (DNA) is a molecule that encodes the genetic instructions used in the development and functioning of all known living organisms and some viruses. Most DNA molecules consists of two complementary nucleic acid strands coiled around each other to form a double helix [25]. The DNA strands are composed of simpler units called nucleotides - guanine (G), adenine (A), thymine (T) or cytosine (C). According to nucleotide pairing rules (A with T and C with G), hydrogen bonds bind the nitrogenous bases of the two separate polynucleotide strands to make double-stranded DNA. Each strand of the original DNA molecule serves as template for the production of the complementary strand during replication. A number of proteins are associated with the replication which helps in terms of the initiation and continuation of DNA synthesis. Most prominently, DNA polymerase synthesizes the new DNA by adding complementary nucleotides to the template strand. (2) In the next step of the flow, the genetic information hard wired into DNA is transcribed into individual transportable cassettes, composed of messenger RNA (mRNA). Thereto RNA polymerases adds com-

plementary RNA nucleotides to subunits of the DNA strands. Each mRNA cassette contains the program for synthesis of a particular protein (or small number of proteins). (3) The basic process of protein production is addition of one amino acid at a time to the end of a protein (Figure 1.3b). This operation is performed by a ribosome. The choice of amino acid type to add is determined by the mRNA molecule. Each amino acid added is matched to a three nucleotide subsequence of the mRNA, called a codon.

This dogma that evolved in 1950s and 1960s was contradicted in 1970 by the discovery of enzymes called reverse transcriptase. This enzyme uses a RNA template to catalyze the synthesis of DNA which is the reverse of the transcription step in the central dogma. These enzymes are encoded and used by reverse-transcribing viruses, such as HIV. These viruses transcribe their RNA genomes into DNA which is then integrated into the host genome and replicated along with it. Reverse transcriptases exhibit high error rates introducing errors at frequencies of one per 1,500 to 30,000 nucleotides during the DNA synthesis and contribute as well to viral diversity.

1.2 Massively parallel sequencing

Massively parallel sequencing encompasses several high-throughput approaches to DNA sequencing where the precise order of nucleotides within a DNA molecule is determined; it is also referred to as next-generation sequencing (NGS) [26]. MPS platforms differ in engineering configurations and sequencing chemistry. They share, however the technical paradigm of sequencing by synthesis mostly for multiple DNA-sequences at the same time. The general workflow to investigate the genetic make up of the virus with MPS is displayed in Figure 1.4 and explained in the subsequent sections.

1.2.1 Library preparation

The library of DNA sequences representing the viral population need to be generated prior to the sequencing. For RNA viruses the viral RNA needs to be converted to DNA. Viral RNA is extracted from plasma samples collected from patients and reverse transcribed. The resulting complementary DNA (cDNA) is used as starting material for the amplification of the drug target region on the viral genome of interest using PCR, a polymerase chain reaction [27]. This is followed by a further enrichment of the region of interest using PCR. PCR is based on the ability of DNA polymerases to replicate DNA fragments (see text box). These polymerases are error-prone similar to the RNA polymerases in the viral replication, but to

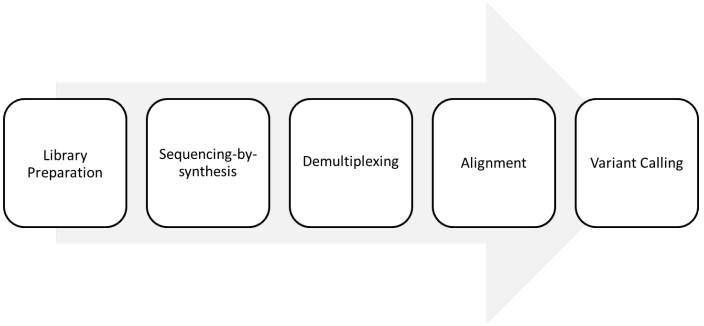


Figure 1.4: The subsequent steps in the massively parallel sequencing process for inferring viral populations are: library preparation, sequencing-by-synthesis, demultiplexing, alignment and variant calling

lesser extent [28]. In addition, the relative frequencies of viral variants can be disturbed by selective amplification bias, especially in low viral load isolates. The DNA sequences containing the drug target region are subsequently fragmented to DNA fragments of a certain length dependent on the MPS platform. The library of these DNA fragments, representing the viral population for a specific drug target, is ready to be sequenced by synthesis. Once these DNA fragments are sequenced they are called reads.

The total number of reads(N), together with the length of these reads (L) and the length of the target region of interest (R) defines the depth (C) of the sequencing process [$C = LN/R$]. The depth, also called coverage, is the average number of reads that represent a given nucleotide position of the viral genome. Sufficient coverage is needed to be able to detect the low frequency variants.

1.2.2 Sequencing-by-synthesis

Sequencing-by-synthesis involves taking a single strand of the DNA fragments to be sequenced and then synthesizing its complementary strand enzymatically. Two different approaches are explained in the following sections.

1.2.2.1 Pyrosequencing

Pyrosequencing platforms, massively parallel sequencing technology commercialized by Roche^{TM} [29], have been the most popular for detection of low-level drug-resistant variants due to their ability to produce long read lengths [30] up to 400 bp. The sequencing is conducted as follows: DNA fragments, representing the viral population, are attached to beads in conditions that favor one fragment per bead. The beads are captured into separate emulsion droplets and PCR amplification occurs within each droplet resulting in beads covered with about 10 million clonal

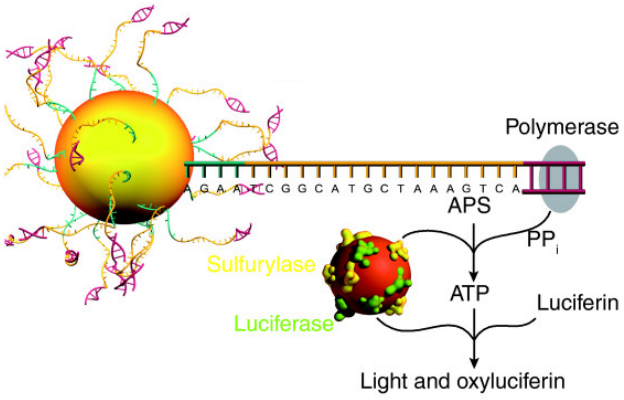


Figure 1.5: In each cycle of the pyrosequencing reaction, one of the four nucleotides is present. Incorporation of the nucleotide in the DNA strand results in the release of PPi that converts APS to ATP. ATP reacts with luciferin to produce light which is captured by lasers. Extrapolation of the intensity signal reveals the number of nucleotides that is build in at this cycle. (source www.454.com)

copies of a single DNA fragment [31]. Each clonally amplified bead is transferred into a single well of a PicoTiterPlate™. All beads are sequenced-by-synthesis simultaneously using a pyrosequencing reaction [32]. A pyrosequencing reaction consist in the cyclic flowing of nucleotide reagents where the repeated flow sequence is T, A, C and G. At each cycle one of the four nucleotides is delivered to the wells in sufficient amounts. Incorporation of a nucleotide (or more than one of the same letter) in the DNA strand results in the release of inorganic pyrophospate (PPi) that converts adenosine 5′ phosphosulfate (APS) to adenosine triphospate (ATP) which react with luciferin to produce oxyluciferin and light (Figure 1.5). The signal intensity at each nucleotide flow, for a particular well, is a proxy for the number of nucleotides - if any - that is incorporated. Quality scores are also derived from these intensities and reflect the probability that the called nucleotide is not an overcall. After each cycle, the excess of nucleotides is washed away, and the next nucleotide in the flow is added. At the end of the sequencing process, reads are obtained for each of the beads containing the nucleotides incorporated over all cycles together with the quality scores for each nucleotide.

The pyrosequencing approach is prone to errors that result from either carry forward errors or incomplete extensions (CAFIE). In the latter case, some DNA fragments on a bead fail to incorporate the nucleotide during the appropriate base flow. These fragments that fail must wait another flow of the other nucleotides before they can continue to sequence and thus those fragments will incorporate out-of-phase with the rest of the fragments. Carry forward errors on the other hand

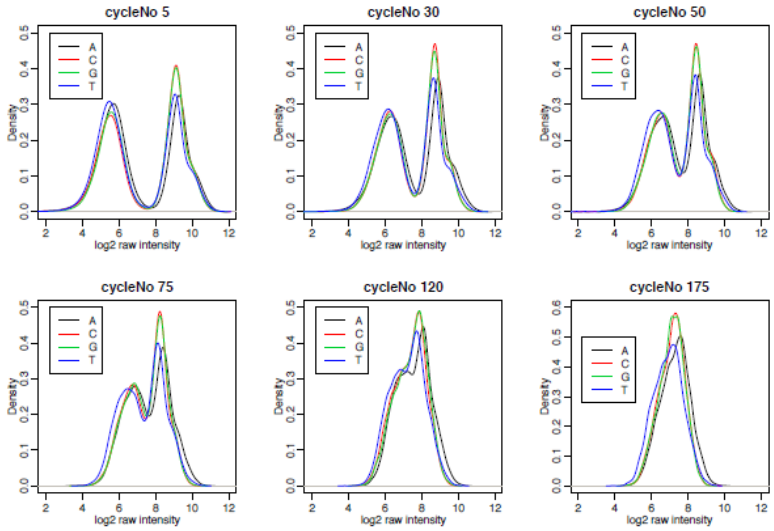


Figure 1.6: Density estimates of the log2 transformed intensities for different cycle numbers and separated per nucleotide. Cycle number 5, for instance, corresponds with the fifth cycle where the nucleotides T, A, C and G were consecutively added. A clear bimodality can be observed corresponding to background and incorporation signal of a particular nucleotide. This bimodality becomes less apparent with increasing cycle number due to phasing [34].

occur when a trace amount of nucleotide remains in a well after the wash, perpetuating premature nucleotide incorporation for specific sequence combinations during the next nucleotide flow. Again it causes those fragments to incorporate nucleotides out-of-phase with the remaining fragments on the bead [33]. These errors accumulate as the sequencing process progresses and near the end more fragments are out of phase. Hence, the difference between background and incorporation signal becomes less apparent with increasing cycle number (Figure 1.6). Additionally the extrapolation of the intensity signal to the number of nucleotides becomes more problematic resulting in homopolymer length inaccuracies in the final sequence.

1.2.2.2 Illumina sequencing technology

Short-read sequencing technology, commercialized by Illumina^{TM} has gradually increased its read lengths [35]. This in combination with a higher sequencing depth at a lower cost per sequenced base makes them an attractive alternative in viral population sequencing. The announcement by Roche to fade out their technology by mid-2016, illustrates the pressing need to focus on alternative technolo-

gies. The performance characteristics of Illumina™ for the characterization of sequence variation in viral population is assessed in the following chapter. The sequencing itself is conducted with the following steps. DNA fragments are spatially segregated on the surface of a glass slide, called flow cell, which consists of 8 lanes for the Genome Analyzer II (one of Illumina's sequencing devices). The DNA fragments are extended to create copies through a series of bridge amplifications resulting in millions of unique clusters (Figure 1.7) [36]. The spatially segregated clusters on the flow cell are sequenced-by-synthesis simultaneously. The complementary DNA strands of the fragments are build up one base at the time making use of fluorescently labeled, reverse terminated nucleotides. Natural competition minimizes incorporation bias since all four reversible terminated nucleotides (dNTPs) are present during each sequencing cycle. After each cycle, the flow cell is imaged in a series of non-overlapping regions, called tiles. The clusters within these tiles, while being excited by laser, generate a quadruplet of intensities, one channel for each nucleotide type. The highest intensity of the four channels determines the nucleotide that is actually incorporated in that cycle. A quality score is derived from each quadruplet of intensities and expresses the probability of calling the wrong nucleotide. The stronger the signal in one of the quadruplets, the higher the probability that it is a correct call. Before the next cycle is started, the fluorescently labeled reversible terminator is cleaved to allow incorporation of the next nucleotide. At the end of the sequencing process, reads with the nucleotide sequences are obtained for each of the clusters together with the quality scores for each base.

Many of the steps of the sequencing process are again error-prone [37]. Overlapping emission spectra of the fluorophores and loss of synchrony of the sequence copies within a cluster can reduce the intensity of the signals which may hamper the correct interpretation of the intensities and which may result in incorrect assignment of bases. The loss of synchrony, similar to 454, introduces particularly errors towards the end of the reads. The marginal distributions for the 4 different bases at three different cycles in the sequencing process are shown in Figure 1.8. It is clear that the incorporation signal of a particular base diminishes towards the end of the reads.

Both the base calls and the quality scores obtained after sequencing are stored in a text-based format, called FASTQ format which normally consist of four lines per sequence [38]. Line 1 begins with a @ character, followed by a sequence identifier which provides the coordinates (lane, tile, x and y-coordinates) of the cluster on the flow cell. Line 2 is the sequence itself. Line 3 begins with a + character and is optionally followed by the same sequence identifier as in line 1. Line 4 represent the quality scores, which are ASCII encoded + 33. This line must contain the same number of symbols as letters in the sequence of line 2. An example is the following:

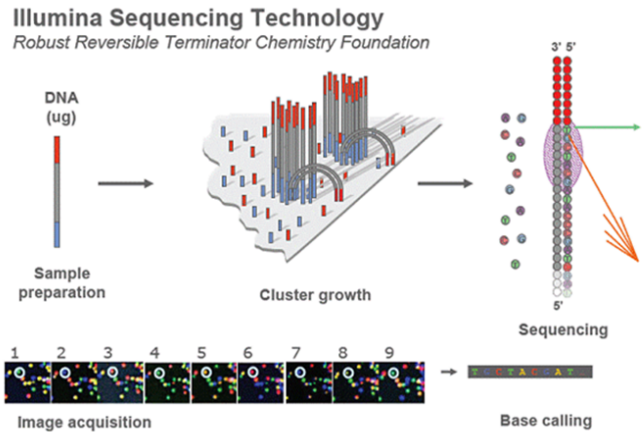


Figure 1.7: Sequencing-by-Synthesis using Illumina's technology. Cluster growth: DNA fragments are copied through a series of bridge amplifications resulting in unique clusters spatially segregated on the flowcell. Sequencing: Each DNA strand is sequenced by synthesis one base at the time (cycle) using fluorescently labeled nucleotides. Image acquisition: After each round of synthesis, clusters are excited by laser emitting the color that identifies the newly added base. Base calling: At the end of the process, reads are obtained containing the bases incorporated in each cycle. (source http://openwetware.org/wiki/BioMicroCenter:Sequencing)

@ HWUSI-EAS1524:12:FC:1:1:18845:1091 1:N:0:
ATGACCCATCAAAAGACTTAATAGCAGAAATACAGAAGCAGGGGCAAGGCC
+
IIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIIIIIIIHIIHIHIHIHFGHF

1.2.3 Demultiplexing and alignment

As massively parallel sequencing are high throughput approaches, several samples can be loaded simultaneously on the PicoTiterPlate^{TM} or on the flow cell for 454 and Illumina respectively. Thereto, the DNA fragments are flanked by sample-specific adapters prior to the sequencing which allows demultiplexing after the sequencing process. After sequencing and demultiplexing, fastq files are obtained for each sample which contains the sequences of the DNA fragments representing the whole viral population. Since, the viral genome was fragmented prior to the sequencing one needs to figure out the corresponding genomic region of each read, a process which is called alignment. The development of alignment algorithms has been successful and many short-read aligners are available to be included in the data analysis pipeline [39].

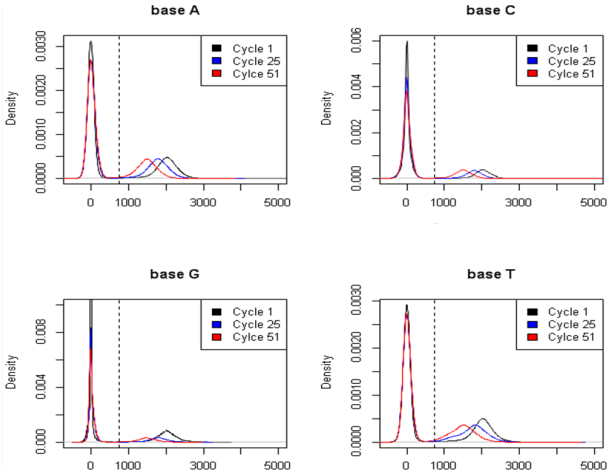


Figure 1.8: Marginal distributions of the intensities for each of the four channels representing one of the 4 nucleotides, for three different positions in the read. The intensity of the signal, corresponding with incorporation of the base, diminishes towards the end of the reads.

1.3 Variant calling

The genetic make up of the DNA fragments at nucleotide level is determined so far together with the location of the fragments in the genome. Hence, the constitution of the viral population for each location in the genome can be investigated. However, the errors introduced during the sequencing process both for Roche™ as Illumina™ may impede a powerful assessment of the low-frequency variants within the viral population as they occur at equal or even higher frequencies than the truly present mutations. Since there are no obvious solutions to reduce the technical noise by further improvements of the technologies, the development of statistical algorithms that can correct the technical noise can be pivotal. Especially, since low-frequency variants are of interest to guide antiretroviral therapies.

Many statistical algorithms have been described in literature to call variants at single-nucleotide level from massively parallel sequencing data. However, most approaches are tailored to call variants in human resequencing projects where variants can be either heterologous (50%) or homologous (100%) [40] since humans are diploid organisms having two copies of the same gene. In viral populations, the variant frequencies cover the full range from 0 to 100% making the error correction more challenging. Some approaches have been made to address this challenge by employing either cut-off based filtering or statistical testing to distinguish true variants from errors. Automated filtering of potential variants based on quality scores are popular, e.g. the VarScan algorithm [41]. Cut-off methods are, however, very

sensitive to parameter choice [42]. Hence, the more promising methods are based on statistical tests where variants are compared to a distribution of errors. Many algorithms are primarily focused on the Roche technology [43–45] as it was the first sequencing technology to become commercially available. These algorithms can not always be transferred to the Illumina technology since this technology reaches higher coverage depth and its quality scores have another interpretation. For Illumina they reflect the probability of a substitution error, while for Roche they reflect the probability of an overcall [46]. The most important methods, applicable for the Illumina technology are the following ones. LoFreq [47] models the error probabilities by using a Poisson-binomial distribution, a generalization of the binomial distribution, where each Bernoulli trial can have a distinct success probability derived from the quality scores. The program V-Phaser 2 [48] recalibrates the quality scores prior to incorporation in the probability model together with co-occurrence of variants within reads. ShORAH [45] is a quality score independent approach where the errors are corrected by applying a model-based probabilistic clustering. The number of clusters is defined using Fisher's exact test to find patterns that occur more frequently than expected by chance.

Most algorithms are focusing on the discovery of single-nucleotide variants. Others, like ShORAH, can be extended to haplotype reconstruction where the co-occurence of variants within the same gene or the same viral genome is investigated. Here, the challenge is to correctly assemble the different variants as they might occur at different DNA fragments. The linkage is done by using overlapping reads. Hence, the assembly of the low-frequency variants [49] is challenging as there might be little overlap. Since low-frequency variants are our main interest, haplotype assembly should be avoided. It is better to restrict to variant calling algorithms within the actual read length. On the other hand, linkage between the nucleotides is lost when calling variants at the single-nucleotide level. However, to have an immediate biological interpretation it is very important to keep the linkage information. Anti(retro)viral drugs target certain proteins which were translated from the viral genome by nucleotide triplets, called codons (Figure 1.3b). Sixty four codons (4^3) translate to twenty different amino acids. Often mutations in the last nucleotide do not result in a change to the amino acid sequence, which is called silent mutation (Figure 1.3c). On the other hand, missense mutations do result in different amino acids. These latter mutations are particularly of interest and can give insight in how the virus builds resistance against anti(retro)viral drugs. Hence, calling variants at the codon level will allow for this immediate biological interpretation at amino acid level. However, none of the existing tools calls variants at the codon level and retrieving linkage between single-nucleotide variants is not straightforward. The research in this dissertation is directed to developing variant callers that act immediately at the codon level and improve biological interpretation considerably.

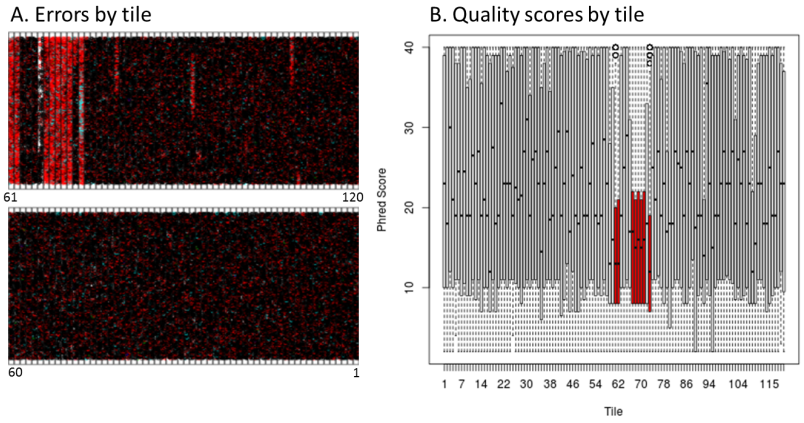


Figure 1.9: a) Clusters of DNA fragments in one lane of the flowcell are plotted based on the x,y-coordinates and divided over 120 tiles. The color indicates the number of errors present in the sequences: black = no error, red = 1 error, blue = 2 errors, ... b) Boxplots of average quality scores of the cluster sequences per tile. Tiles where more errors occur during the sequencing process have on average lower quality scores.

1.4 Challenges

A major challenge within virology applications remain the detection of low frequency variants as the errors introduced in the different steps of the sequencing process may occur at equal or even higher frequency. The sequencing quality scores of Illumina can help here, because they reflect the probability of a substitution error. In Figure 1.9a an image is provided of the different clusters of DNA fragments in one lane of the flow cell distributed over 120 tiles (based on the x,y-coordinates of the clusters). In Figure 1.9b the distribution of average quality scores of these sequences per tile are plotted. Comparison of these two figures reveals that tiles with more errors in the sequences, have on average lower quality scores. Hence, these quality scores will be used in a filtering approach where the number of false-positive findings are reduced by filtering out the low quality variants. The cut-offs are defined based on the data to account for differences in quality between the sequencing runs. The algorithm is described in chapter 3.

The quality scores, however, do not always reflect the true error probabilities. In Figure 1.10 the quality scores are plotted against the theoretical error probabilities (green) and the observed error probabilities (orange). The Phred quality scores (Q) are logarithmically related to the error probability (E)

$$Q = -10logE \tag{1.1}$$

The observed error probabilities are calculated using a dataset with known vari-

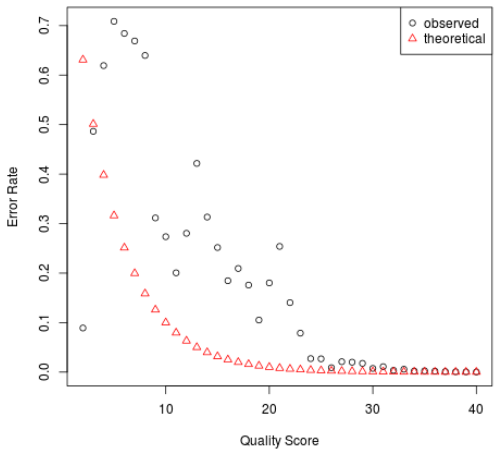


Figure 1.10: Error probability as a function of the quality scores. The observed error probabilities are plotted together with the theoretical error probabilities in red.

ation. For each quality scores, the number of false positives are compared with the total number of nucleotides that have the same quality score. It is clear that the quality scores underestimate the true error probabilities. Hence, additional metrics might be preferred to distinguish technical noise from low-frequency variants. The Illumina technology calls the base that corresponds with the highest intensity among four fluorescence channels and the quality scores are derived from these intensities. Extra information which can be exploited are the second best base calls, the bases corresponding with the second highest intensity. Substitution errors can often be corrected by these second best base calls [50]. This is investigated in chapter 4.

1.5 Outline and contributions

The research project presented in this dissertation was granted by IWT, Baekeland 100679, at the end of 2010 and started officially in January 2011. Baekeland mandates are projects assigned to consortia involving a Flemish company and university, which is Janssen Pharmaceuticals and Ghent University respectively. The purpose of such mandates is to support basic research that has clear economic objectives and offers added value to the company involved in the project. The first challenge in such projects is to get the university and the company aligned. Hence, most part of the first year was spend to understand the need of the company and to learn the languages of the other people in these multidisciplinary teams. To

facilitate this, I was mostly located at Janssen Pharmaceuticals in Beerse where I could interact with the non-clinical statistics group, as well as with their clients in discovery sciences.

1.5.1 Outline

The outcome of the project is displayed in the following chapters using the three papers which are published or under revision.

1. Thys K., Verhasselt P., Reumers J., Verbist B.M.P, Maes B., Aerssens J. Performance assessment of the Illumina massively parallel sequencing platform for deep sequencing analysis of viral minority variants. *under revision at Journal of Virological Methods*.

2. Verbist B.M.P., Thys K., Reumers J., Talloen W., Aerssens J., Clement L., Thas O. VirVarSeq: a low frequency Virus Variant detection pipeline for Illumina Sequencing using adaptive base-calling accuracy filtering. 2014 *Bioinformatics* doi: 10.1093/bioinformatics/btu587.

3. Verbist B.M.P., Clement L., Reumers J., Thys K., Vapirev A., Talloen W., Wetzels Y., Meys J., Aerssens J., Bijnens L., Thas O. ViVaMBC: estimating Viral sequence Variation in complex populations from Illumina deep-sequencing data using Model-Based Clustering. *under revision at BMC Bioinformatics*.

The two developed variant callers VirVarSeq and ViVaMBC, described in chapter 3 and 4, are compared in the last chapter followed by a discussion and the valorization of the project. The papers itself resulted from the work performed in multi-disciplinary teams which is typical in these Baekeland mandates. In the following I will briefly specify my contributions in each of these papers.

1.5.2 Contributions

For the first paper (Chapter 2), support was given to the first author, Thys K., for the assessment of the reporting limit which could be reached by Illumina since the technology was not yet widely applied in the field of viral quasispecies detection. Further some support was given in writing the discussion of the final paper. Involvement in this project was necessary for understanding the current practice in reporting variants within Janssen Pharmaceuticals and to learn the challenges involved.

Building further on the way of reporting I developed Q-cpileup which is described in the second paper (Chapter 3). Q-cpileup builds further on existing code of McLachlan and was made available from R by Joris Meys. Q-cpileup eventually

ended up as a component of the automated pipeline VirVarSeq. Of course a thorough testing on multiple samples was needed and was performed by myself. The pipeline itself was made operational in collaboration with Reumers J. and Wetzels Y. A users guide was written together with Wetzels Y. At the time of writing this thesis, VirVarSeq is used within Janssen Pharmaceuticals to support some of their clinical studies.

In the third paper (chapter 4), we went one step further and developed a model based clustering approach. The method was developed in collaboration with my academic promoters of Ghent University. It was implemented by myself with some R help from Joris Meys. Again some thorough testing was needed. Help was provided here by Yves Wetzels to run the code in the amazon elastic compute cloud. Intermediate reporting to check with reality was driven by myself. The paper and code were written by me and code was parallelized by Alexander Vapirev to improve performance.

2

Performance Assessment of Illumina

K. Thys, P. Verhasselt, J. Reumers, B.MP Verbist, B. Maes, J. Aerssens

Performance assessment of the Illumina massively parallel sequencing platform for deep sequencing analysis of viral minority variants. (2014), under revision at Journal of Virological Methods .

Abstract *Massively parallel sequencing (MPS) technology has opened new avenues to study viral dynamics and treatment-induced resistance mechanisms of infections such as human immunodeficiency virus (HIV) and hepatitis C virus (HCV). Whereas the Roche/454 platform has been widely used for the detection of low-frequent drug resistant variants, more recently developed short-read MPS technologies have the advantage of delivering a higher sequencing depth at a lower cost per sequenced base. This study assesses the performance characteristics of Illumina MPS technology for the characterization of genetic variability in viral populations by deep sequencing. The reported results from MPS experiments comprising HIV and HCV plasmids demonstrate that a 0.5-1% lower limit of detection can be readily achieved with Illumina MPS while retaining good accuracy also at low frequencies. Deep sequencing of a set of clinical samples (12 HIV and 9 HCV patients), designed at a similar budget for both MPS platforms, revealed a comparable sensitivity/lower limit of detection for Illumina and Roche/454. Finally, this study shows the possibility to apply Illumina's paired-end sequencing as a strategy to assess linkage between different mutations identified in individual vi-*

ral subspecies. These results support the use of Illumina as another MPS platform of choice for deep sequencing of viral minority species.

2.1 Introduction

RNA viruses such as human immunodeficiency virus (HIV1) and hepatitis C (HCV) are genetically diverse due to high replication rates, relatively small genomes and error prone polymerases [8, 69–71]. Even within the host, RNA viruses exist as complex populations composed of several closely related subgroups, so-called quasi-species [10, 11]. Furthermore, host factors and immune responses apply a selection pressure that adds to diversification of the quasi-species. A highly heterogeneous mixture of viral genomes enables the virus to adapt rapidly to changing environments and develop resistance to antiviral therapy. For example, HIV reverse transcriptase is an important target for antiviral therapy but can harbor several drug-resistant mutations, predominantly occurring within the first 350 amino acids [72]. Similarly, the NS3 genomic region in HCV is encoding for its protease and is the molecular target for the recently approved direct antiviral drugs Telaprevir and Boceprevir [73]. Several mutations, predominantly found in the first 181 amino acids of NS3, have already been associated with resistance to these drugs [74]. Sanger sequencing of PCR-generated amplicons of these target regions of interest has been applied for many years as the "gold standard" to detect drug-resistant variants in clinical samples. Although this "population sequencing" method reliably identifies the major mutations, it fails to detect viral subspecies present at frequencies below 20-30% in a viral population [75]. Several studies however emphasize the clinical relevance of low frequency drug resistant variants [16, 76, 77] entailing the need of a more in depth characterization of sequence variation. Massively parallel sequencing (MPS) technologies enable the deep sequencing of viral populations with much greater sensitivity. Roche's 454 technology was the first available deep sequencing platform in the field and has been used widely to detect low abundant drug-resistant variants in viral applications, especially also because the relatively long read lengths allow to identify linked mutations present on the same viral genome [78–80]. Today, however, alternative massively parallel short-read sequencing technologies such as the Illumina platform have become available that can provide higher sequencing depth at a lower cost per base. The recent announcement by Roche to retract the 454 sequencing technology from the market soon, illustrates the pressing need to evaluate alternative technologies. The main objective of this study was to challenge Illumina's GAIIx sequencing technology to at least the same sensitivity as generally accepted for 454, where variants down to 1% can be detected [29, 49, 81]. Interestingly, whereas novel technology developments are continuously further improving the technical sensitivity to detect low frequency viral variants, recent reports suggests

that the clinical impact of HIV minority species becomes potentially manifest only for variants present above 1% [20]. Hence, especially the window between 1% and 20% might be of particular clinical interest to determine minor viral variants with high accuracy. Whereas the higher sequencing coverage generated by the Illumina platform would theoretically also result in a higher sensitivity as compared to the Roche/454 technology, it is obvious that the lower limit of detection for minority variants is bound by error rates that may originate from diverse steps during both the library preparation (typically including PCR amplification) [83] and the sequencing process itself [84, 85]. Despite reduced error rates as a result of optimized wet lab protocols [86–88], the differentiation of technical errors from true low frequency variants remains one of the major challenges in deep sequencing. Therefore different aspects were assessed that might contribute to the accuracy performance of the Illumina technology in comparison to the established Roche/454 technology for deep sequencing applications in virology. In addition, this study demonstrates the possibility to derive viral mutation linkage information from the short read sequences generated by Illumina technology. When multiple clinically relevant mutations are identified in a viral sample, it can be important to know whether these mutations are present on the same or different viral genomes. Until recently, this information could only be gathered through labor-intensive cloning by limited dilution and subsequent Sanger sequencing of many individual viruses from a clinical viral population. Today, MPS platforms can produce thousands of clonal sequences in a rapid and cost effective way. Linkage of mutations and haplotype reconstruction can be accomplished using long read MPS technologies such as Roche/454, albeit constrained by the physical read length (approximately 400 base pairs) [89]. Haplotype reconstruction is particularly challenging for short-read technologies [90], but Illumina's paired-end sequencing approach enables linkage analysis beyond the limit of the read length itself. Whereas this approach was originally developed to improve read alignment, this study demonstrates its concomitant value for linkage analysis.

2.2 Materials and methods

2.2.1 Plasmid samples

Two different HCV plasmids were used, each comprising the viral NS3-4A fragment. Site-directed mutations have been introduced into the con1b replicon plasmid pFK_i341_PI Luc_NS3-3'_ET (wild type) as described earlier [91,92]. These plasmids (wild type and mutant) differ only in two single nucleotides from each other, as confirmed by Sanger sequencing. For MPS experiments, the fragment encompassing the NS3-4A region (2.4 kb) was PCR-amplified using region-specific primers [80] and Phusion hot start high-fidelity DNA polymerase, Finnzymes Oy,

Espoo, Finland) according to manufacturer's instructions. PCR reactions were done in triplicate and subsequently pooled to reduce possible PCR bias. Similarly, four HIV plasmids comprising the reverse transcriptase (RT) gene in a pGEM-3zf background were used in the study. In an experiment that focused on the evaluation of errors introduced during the MPS sequencing process, the whole plasmid was sequenced thereby avoiding potential errors that might otherwise be introduced during target amplification.

2.2.2 Clinical samples

Viral RNA was extracted from plasma samples collected from HCV and HIV patients using the automated NucliSENS® easyMAG® system (bioMérieux). The viral RNA (16 μl) from HCV patients was reverse transcribed using random hexamers (Invitrogen, Carlsbad, CA) and Accuscript™ high fidelity reverse transcriptase (Agilent Technologies, Santa Clara, CA). The resulting cDNA (2 μl) was used as starting material for amplification of a fragment encompassing the HCV NS3-4A region (2.4 kb) by two-round nested PCR using gene-specific primers [93] and KOD DNA polymerase (Novagen, Madison, WI). For HIV samples, viral RNA (10 μl) was reverse transcribed and cDNA was amplified in a one-step PCR using gene-specific primers [94] (Super-Script™ III One-Step RT-PCR System with Platinum® Taq DNA Polymerase (Life Technologies, Carlsbad, CA). A 1.9 kb amplicon encompassing the HIV protease (PR) - RT region was subsequently amplified in a nested PCR using the Expand high fidelity PCR system (Roche Applied Science, Pesberg, Germany) or Phusion hot start high-fidelity DNA polymerase (Finnzymes Oy). For both HIV and HCV, PCR reactions were done in sevenfold and subsequently pooled to reduce possible PCR bias.

2.2.3 Amplicon purification and quantification

Prior to fragmentation and deep sequencing, all amplicons, derived from either plasmids or clinical samples, were purified using the QIAquick PCR purification kit (Qiagen, Venlo, Netherlands) and QIAquick gel extraction kit (Qiagen) or Agencourt Ampure XP (Beckman Coulter Genomics, Danvers, MA). Samples were quantified using the Quant-iT™ PicoGreen® dsDNA kit (Life Technologies).

2.2.4 Illumina deep sequencing

Following Illumina's standard protocols, the DNA (0.1-0.5 μg) was fragmented to an average length of 200 bp using the Covaris® E210 system (Covaris, Woburn, MA). The ends of the fragmented DNA were repaired, adenylated and Illumina compatible adaptors (Index PE Adaptor Oligo Mix (Illumina, San Diego, CA)

or barcode-included adaptors from NEXTflex™ DNA Barcodes (Bioo Scientific, Austin, TX)) were ligated using the SPRIworks Fragment Library System I (Beckman Coulter Genomics). In case when Index PE Adaptor Oligo Mix adaptors were used, fragments were indexed using Illumina compatible barcodes by the Multiplexing Sample preparation Oligonucleotide Kit (Illumina). Next, the library was enriched during 12 cycles of PCR. Enriched fragments were visualized on a Bioanalyzer (Agilent Technologies) for quality control and quantification. Next, samples were pooled according to the specific experimental setup, prior to applying on the Illumina cluster station for cluster generation using the TruSeq PE Cluster Kit v2 (Illumina). A multiplexed paired-end sequencing run of 147 cycles was executed using the TruSeq SBS Kit v5 (Illumina) on the Genome Analyzer IIx (GAIIx) (Illumina).

2.2.5 454 deep sequencing

Deep sequencing by 454 was performed as described earlier [94]. In short, DNA was fragmented to an average length of 500 bp using the Covaris® E210 system (Covaris). Using the SPRIworks Fragment Library System II (Beckman Coulter Genomics), the ends of the fragmented DNA were repaired and adenylated and 454 sequencing adaptors were ligated to allow for sample multiplexing in sequencing lanes on the Roche GS FLX instrument. Fragments were indexed using the GS FLX Titanium Rapid Library MID Adaptors kit (Roche). Next, samples were pooled according to the experimental setup. Clonal amplification was performed by emulsion PCR (GS FLX Titanium emPCR Kit, Roche). Finally, samples were sequenced using the GS FLX Titanium Sequencing Kit (Roche) on a GS FLX instrument (Roche).

2.2.6 Initial MPS data processing

First, a consensus sequence representing the majority of the underlying viral population was derived for each of the samples. This consensus sequence was either obtained through independent Sanger population sequencing, or derived from the available MPS reads after mapping against a universal reference sequence (HIV or HCV). Next, all individual MPS sequence reads were mapped for each sample against their own consensus sequence, using CLCBio Workbench (CLCBio, Aarhus, Denmark). Mapping parameters were set to favor single nucleotide mismatches over single nucleotide insertions or deletions considering the viral coding background. A similarity of 80% with the reference was decisive for mapping 50% of the read length allowing low quality ends of the read to be trimmed for further analysis. Minority variant analysis focused on a subset of the alignment, namely the 1050 bp region corresponding to HIV reverse transcriptase amino acids 1-350 and the 543 bp region encoding the HCV protease amino acids 1 to 181. Codon

variants were determined per amino acid position of the region of interest (HIV: amino acids 1-350 of RT gene; HCV: amino acids 1-181 of NS3 gene) and their relative frequencies were calculated as the proportion of reads containing a specific codon divided by the coverage at that position. Relative frequencies of codon variants were calculated separately for each sequencing direction (forward and reverse); the lowest observed frequency (either forward or reverse) per codon variant was reported. As demonstrated earlier, the insertion of errors during high throughput sequencing is highly dependent on the direction of the sequence read hence, bi-directionality is often used as a factor for error correction/filtering [84, 95, 96]. By means of additional error filtering for Illumina sequencing, codons with at least one nucleotide with a quality value lower than 30 (representing an error rate of 1 in 1000 [56] were not taken into account.

2.2.7 Coverage depth analysis

In order to determine the sequencing depth needed to maintain accurate quantitative assessment of minor variant frequencies, 4 mixtures containing different ratios (1:200, 1:100, 1:50, and 1:10) of wild type and mutant HCV plasmids harboring two codon variant at NS3 position 36 and 155 were deep sequenced using the Illumina platform. Random subsets of sequence reads (0.1, 0.2, 0.5, 1 and 2 million reads) were generated from the total dataset (average 7.4 million reads per sample) to simulate different levels of coverage ($\pm$ 1,000x, 2,000x, 5,000x, 10,000x, and 20,000x, respectively; performed in triplicate). For each coverage depth, frequencies of all codon variants (errors included) at NS3 positions 36 and 155 were calculated.

2.2.8 Linkage analysis

The degree of linkage between observed mutations at amino acid positions 101 and 138 in the HIV-RT gene, located 114 bp apart - and thus only identifiable on different individual reads (only 70 bp read length in the Illumina run) – was calculated. DNA was fragmented in approximately 200 bp fragments during sample preparation and subsequently paired-end sequenced. The sequence information of both ends of the same DNA fragment is available and can be coupled through positional information (XY coordinates of the Illumina flow cell) hence linked mutations can be derived for individual viral subspecies. From this information, percentages of linked mutations in fragments containing all possible combinations of codon variants were derived.

Expected freq (%)	NS3_36 Observed freq (%)	NS3_155 Observed freq (%)
10	11.28	10.35
2	2.37	2.14
1	0.91	0.83
0.5	0.44	0.32
0	0	0

Table 2.1: Expected and observed frequencies of a spiked-in HCV-NS3/4A plasmid harboring two codon variants⋆ as revealed by Illumina deep sequencing. ⋆Mutant HCV NS3/4A plasmid, carrying two single codon variants (codons 36 and 155) was spiked-in at different ratios (1:10, 1:50, 1:100, 1:200 and 0:1) versus the wild-type plasmid.

2.3 Results

2.3.1 Accuracy of Illumina deep sequencing

Deep sequencing holds the promise to quantitatively assess the presence of minority variants in viral populations, which cannot be reliably determined by Sanger sequencing. The accuracy of deep sequencing on the Illumina GAIIx platform was evaluated by spiking a HCV plasmid carrying two codon variants in the NS3 region at different ratio's (1:10, 1:50, 1:100, 1:200, and 0:1) into wild type HCV plasmid, corresponding to variant frequencies between 10% and 0%. Targeted deep sequencing of the fragment encoding amino acids 1-181 of HCV NS3 revealed an average coverage depth of 64,877x, ranging from 32,574x to 114,925x, for the 5 plasmid mixtures. Comparison of the expected and observed frequencies for all spiked-in variants demonstrates the highly accurate quantification that can be achieved with ultra-deep sequencing (Table 2.1). Although theoretically such high coverage would allow the detection of variants far below 0.5%, the presence of errors introduced by wet lab preparation methods and sequencing complicates the reliable detection of low frequency variants. For the plasmid mixtures, each codon differing from the consensus sequence other than the spiked-in variant is considered to be an error, although one cannot exclude that some of the observed sequence differences are real and could be attributed to errors introduced during plasmid preparation. Figure 1 shows the observed frequency of the spiked-in mutations in the background of all other variants observed in the HCV NS3 region. This analysis confirms the reliable detection of variants spiked-in at frequencies above 0.5% (Figure 2.1A), but also demonstrates a high abundance of low-frequency errors that may obscure the identification of true variants present at low frequency (Figure 2.1B).

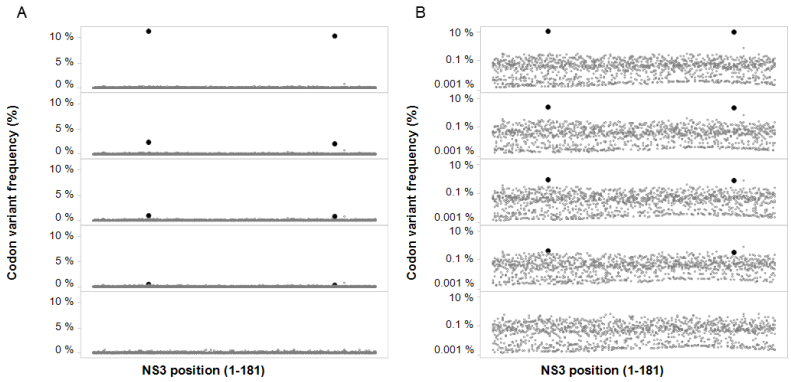


Figure 2.1: Frequencies of codon variants (in percentage) per position in HCV-NS3 represented in (A) absolute and (B) logarithmic scales. HCV plasmid carrying two codon variants was spiked at different ratios (1:10, 1:50, 1:100, 1:200, 0:1) in wild-type plasmid. Each dot represents a single codon variant, with the spiked-in variants depicted in black color and all other detected codon variants, considered to be errors, in gray. A codon variant was defined as any codon sequence that is different from the consensus codon sequence.

2.3.2 False discovery rate of Illumina deep sequencing

Next, the false discovery rate (FDR) of viral minor variants was assessed through deep sequencing of wild type plasmids. Four plasmids carrying HIV-RT amino acids 1-350 and 2 plasmids carrying HCV NS3 amino acids 1-181 were analyzed using the Illumina GAIIx sequencing platform. Average read numbers of approximately 2.6 and 6.9 million yielded an average coverage of 13,739 and 60,137 reads per position for the HIV and HCV plasmids, respectively. Errors from the reference plasmid sequence were observed in 269,193 out of 57,631,296 sequenced nucleotides (0.47%) for the 4 HIV plasmids, and in 172,295 / 65,262,693 sequenced nucleotides (0.26%) for the 2 HCV plasmids. At codon level, this implied 144,610 / 19,210,432 (0.75%) incorrect codons for the 4 HIV plasmids, and 171,726 / 21,754,231 (0.79%) incorrect codons for the 2 HCV plasmids. The observed error rate at codon level is higher as one nucleotide change in the triplet is enough to change the codon. Because sequencing errors are not randomly distributed and can vary considerably per position [84], the application of the average error level as a fixed cut-off appears not sufficiently stringent to accurately discriminate true variants from the false positive background noise. Whereas in all examined plasmids, the vast majority of observed variants were found at low frequencies (P90 at 0.2%-0.3%), few of them reached frequencies up to 1% (Figure 2.2). The false discovery rate for the HIV and HCV plasmids was calculated at increasing cut-off values (Figure 2.3), demonstrating that the FDR for the detection of minor vari-

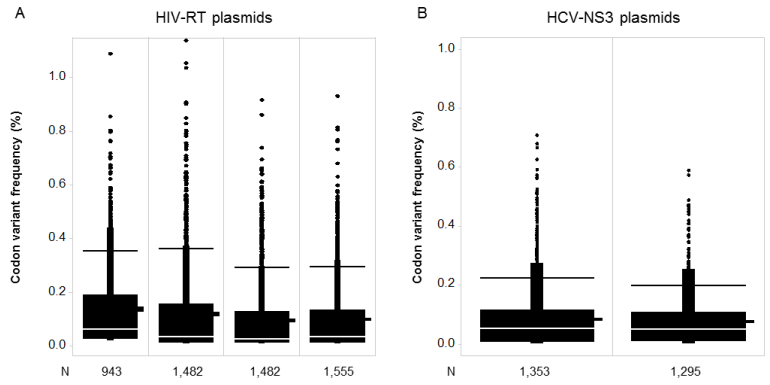


Figure 2.2: Box plots representing relative frequencies of variants in 6 clonal samples, comprising (A) HIV-RT plasmids (n=4) or (B) HCV-NS3 plasmids (n=2), observed by Illumina sequencing. Each dot represents a single variant (codon sequence different from the consensus codon sequence). N indicates the total amount of false positive variants. The black line represents the 90th percentile range (P90); the white line depicts the median frequency of minority codon variants.

ants converges to zero for all plasmids well below a cut-off value of 1%. It should be noted that the above analysis based on plasmid sequences might underestimate the FDR in clinical viral RNA samples, because plasmid DNA samples do not undergo a reverse-transcription PCR step during library preparation. Since this latter step could be an important source of sequencing errors, AccuScript high-fidelity reverse transcriptase (Stratagene) which displays a threefold higher fidelity (error rate 2x 10-5) than commonly used MMLV reverse transcriptase [136] was utilized for the processing of clinical viral RNA samples.

2.3.3 Coverage depth analysis

Given the large discrepancy between coverage depths used in typical deep sequencing experiments using Illumina and Roche/454 technologies, an experiment was designed aiming to determine the sequencing depth needed to maintain accurate quantitative assessment of minor variant frequencies. Sample mixtures containing different ratios (1:10 up to 1:200) of HCV plasmids harboring two codon variants (at NS3 position 36 (GTC→ATG) and 155 (CGG→AAA) versus wild type) were sequenced using the Illumina GAIIx sequencing platform. On average, sequencing yielded 7.4 million reads per sample mixture. Random subsets of sequence reads (0.1, 0.2, 0.5, 1 and 2 million reads) were generated in triplicate from the total dataset in order to mimic different levels of coverage (1,000x, 2,000x, 5,000x, 10,000x, and 20,000x, respectively). For each sequencing depth, all codon variants identified at NS3 position 36 were reported, errors included

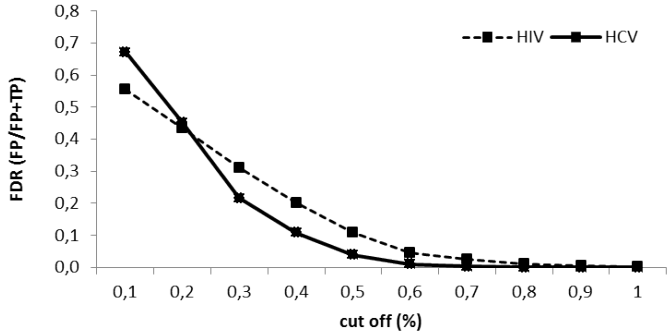


Figure 2.3: False discovery rate (FDR) in function of cut-off value for HIV-RT (n=4) and HCV-NS3 (n=2) plasmids. FDR was calculated as the amount of false positives (FP) divided by the total amount of positives (true and false). True positives (TP) were defined as consensus codons covering the region of interest (350 codons per HIV plasmid and 181 codons per HCV plasmid).

(Figure 2.4). This visualization enables to evaluate both the accurate detection of the NS3 variant as well as the quantitative assessment of codon frequencies. For samples comprising the NS3 variant spiked-in at 10% or 2%, the correct variant could be reliably detected above the error frequencies for all coverage depths (Figure 2.4A and B). Observed variant frequencies ranged from 9.01% to 11.01% and from 1.85% to 2.65%, respectively, well above the maximum error frequencies (0.13% to 0.55%) observed at this position. The standard deviation of the observed frequencies decreased at higher coverage, indicating more **precise** variant frequency estimations at higher sequencing depth. Although variant frequency was generally underestimated for plasmid mixtures with 1% spiked-in variants (frequency ranging from 0.24% to 1.07%), the estimated frequency improved with increasing coverage (Figure 2.4C). At the lowest coverage depth tested ($\pm$1,000 reads), it was impossible to distinguish the spiked-in variant from the errors in two out of three replicates. Finally, NS3 variants spiked-in at 0.5% could only be distinguished from errors at the highest tested coverage's (Figure 2.4D).

2.3.4 Illumina versus 454 deep sequencing

In order to assess the lower limit of detection in clinical samples, 12 isolates of HIV patients and 9 isolates of HCV patients were analyzed in parallel with both Illumina and Roche/454 sequencing platforms. For HIV isolates, the HIV PR-RT (1.9 kb) region harboring the first 350 amino acids of HIV-RT were amplified and sequenced, while a 2.4 kb amplicon (encoding HCV NS3-4A) was designed

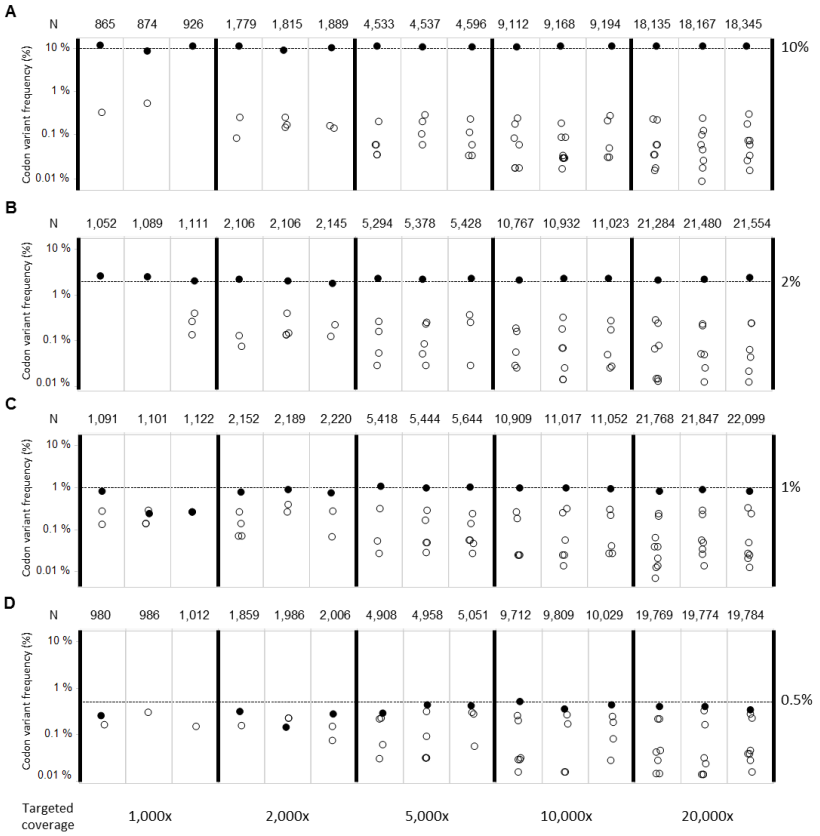


Figure 2.4: Accuracy of codon variant frequency estimation in function of sequencing depth. An HCV plasmid carrying two codon variants (at position 36 and 155 of NS3) was spiked at different ratios (1:10, 1:50, 1:100, 1:200) in wild-type plasmid. The individual graphs show the frequency of all observed codon variants at amino acid position 36 of NS3 for different levels of sequencing coverage. The coverage (N) is the result of random sampling (S) of sequence reads from the original dataset (1, 2, 5, 10 or 20 x105 reads). Samplings were performed in three-fold. Each circle represents one observed codon variant. The filled circle represents the spiked codon variant (ATG) whereas open circles represent all other codon variants which are considered to be errors. The dotted line represents the expected frequency of the spiked variant.

to inspect the first 181 amino acids of the HCV NS3 protein. Illumina sequencing yielded on average 3.8 million reads for the HIV-RT amplicon and 1.4 million reads for the HCV-NS3 amplicon, resulting in an average coverage of 99,660 and 19,718, respectively. At a comparable cost, 454 sequencing resulted in average read numbers of 53,899 and 15,497 and an average coverage of 8,613 and 1,347 reads per position for HIV and HCV, respectively. Minor variant frequencies observed by both technologies were compared for all clinical isolates. Above the conservative 1% cut-off, variant frequencies observed in the two technologies were highly correlated ($r^2 = 0.97, n = 835, p < 0.0001$ for the HIV isolates; $r^2 = 0.96, n = 266, p < 0.0001$ for the HCV isolates) (Figure 2.5A and B). Below this 1% frequency threshold, bigger discrepancies between the frequencies are observed. Variants detected by either of both platforms -above the pre-defined 1% cut-off- were considered as possible false positives/negatives. To account for small bias in frequency estimation, all variants that were detected at frequencies less than 0.5% in one platform and above 1% in the other were investigated further. For HIV samples, 1 variant was identified with Illumina that was present below 0.5% on Roche/454 whereas 38 variants detected with Roche/454 were not observed at frequencies above 0.5% with Illumina. All the latter variants were located in homopolymeric stretches of nucleotides and were therefore considered Roche/454 false positives. For the HCV samples, an equal amount of variants were seen with Illumina (32) that were present below 0.5% with Roche/454 and vice versa (29). These errors occurred at maximum frequencies of 3.3% and 2.0% respectively. No trend in the nature of mutations could be appointed for HCV. However, due to the low coverage of the 454 experiment a bias in accurate estimation of frequencies cannot be excluded which can lead to both over- and under-estimation of 454 variants compared to Illumina.

2.3.5 Linkage analysis

In addition and complementary to a reliable detection and quantitative assessment of variant frequencies, the study of viral populations in clinical isolates often involves the study of linkage between resistance-associated mutations. Given the relatively short read length of Illumina sequencing technology, it is often not considered as an appropriate technology to perform such analysis. This study demonstrates that it is feasible to investigate linkage using the Illumina sequencing platform, given the appropriate experimental conditions are used. The complete HIV-RT region (1.9 kb) was amplified from 9 clinical isolates harboring IAS-USA resistance-associated mutations (RAM) K101E and/or E138K (according to Sanger population sequencing). Subsequent library preparation for the Illumina GAIIx involved a fragmentation step revealing average fragment sizes of 200 bp, which should be large enough to contain both RAM positions in sin-

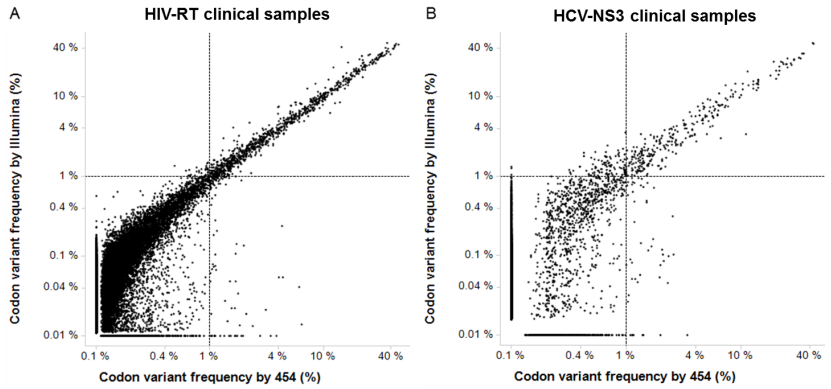


Figure 2.5: Scatterplot representing relative frequencies of codon variants detected by Illumina versus Rochde/454 sequencing of the HIV-RT region of 12 clinical samples (A) and HCV-NS3 region of 9 clinical samples (B). Codon variants are defined as codons different from the consensus codon of the sample. Each dot represents one codon variant of one sample. Frequencies are plotted in log-transformed scale. The dotted lines indicate the 1% cut-off level (both for Illumina and Roche/454), representing the analytical sensitivity typically applied for Roche/454 sequencing.

gle DNA fragments. All isolates were sequenced with 70 bp paired-end reads at the Illumina GAIIx sequencing, resulting in 7.5 million reads (on average) per sample. Following the mapping of the reads (yielding an average coverage of 213,355x and 251,117x at positions 101 and 138, respectively), variant frequencies of K101E and E138K were calculated. A strong inverse correlation ($r^2 = 0.989, p < 0.0001$) was observed between the frequencies of both mutations (Figure 2.6), suggesting mutual exclusivity of both mutations. Next, linkage analysis was performed to assess whether both mutations could co-exist in one viral genome. The dataset of sequencing reads was filtered to contain only paired reads (sequences from both ends of the same fragment) spanning both positions 101 and 138, revealing on average (for the 9 isolates) 45,040 paired reads covering both positions of interest. The majority of paired reads from all 9 isolates harbored only one of the 2 mutations (E138K and K101E) in combination with wild-type (49% and 40%, respectively). Five percent of all paired reads were wild type at both positions and 2% carried other variants, leaving only 4% of all the reads (from all isolates) harboring both mutations (Table 2.2). The probability of observing mutant E138K independent of the other mutation is 53%, while it is 44% for mutant K101E. This implies that the probability of observing both observations together, if they are independent from each other, is 23%. Hence, the observed 4% is far below and hence the mutations E138K and K101E are mutually exclusive and appear rarely on the same viral genome.

Sample	K101K/ E138E	K101K/ E138K	K101K/ E138⋆	K101E/ E138E	K101E/ E138K	K101E/ E138⋆	K101⋆/ E138E	K101⋆/ E138K	K101⋆/ E138⋆	Total
1	172	215	4	41,609	508	303	357	6	2	43,176
	(<1%)	(<1%)	(<1%)	(96%)	(1%)	(1%)	(1%)	(<1%)	(<1%)	
2	5,021	6,626	80	30,168	1,632	126	1,112	103	6	44,874
	(11%)	(15%)	(<1%)	(67%)	(4%)	(<1%)	(2%)	(<1%)	(<1%)	
3	2,947	9,042	6	31,819	2,656	26	91	34	0	46,621
	(6%)	(19%)	(<1%)	(68%)	(6%)	(<1%)	(<1%)	(<1%)	(0%)	
4	2,300	7,122	131	17,628	2,516	77	54	30	0	29,858
	(8%)	(24%)	(<1%)	(59%)	(8%)	(<1%)	(<1%)	(<1%)	(0%)	
5	2,880	8,311	41	19,033	3,663	114	80	40	0	34,162
	(8)	(24%)	(<1%)	(56%)	(11%)	(<1%)	(<1%)	(<1%)	(0%)	
6	2,829	19,537	80	12,657	2,593	65	48	82	0	37,891
	(7%)	(52%)	(<1%)	(33%)	(7%)	(<1%)	(<1%)	(<1%)	(0%)	
7	1,315	37,016	54	4,651	1,201	5	16	77	0	44,335
	(3%)	(83%)	(<1%)	(10)	(3%)	(<1%)	(<1%)	(<1%)	(0%)	
8	1,571	52,116	3,736	2,587	1,947	341	15	144	8	62,465
	(3%)	(83%)	(6%)	(4%)	(3%)	(1%)	(<1%)	(<1%)	(<1%)	
9	1,243	58,599	460	833	633	6	7	193	1	61,975
	(2%)	(95%)	(1%)	(1%)	(1%)	(<1%)	(<1%)	(<1%)	(0%)	
Total	20,278	198,584	4,592	160,985	17,349	1,063	1,780	709	17	405,357
	(5%)	(49%)	(1%)	(40%)	(4%)	(<1%)	(<1%)	(<1%)	(<1%)	(100%)

Table 2.2: Clonal distributions of observed variants at codon positions 101 and 138 in the HIV RT gene. The paired-end Illumina sequencing method enabled to calculate counts (and row percentages) of paired sequence fragments for each of the different possible combinations of codon variants at position 101 and 138 (residing on the different sequence fragments). Observed amino acid variants are wild type (K101K and E138E), resistance-associated mutations (K101E and/or E138K) or any other (designated as K101⋆ or E138 ⋆).

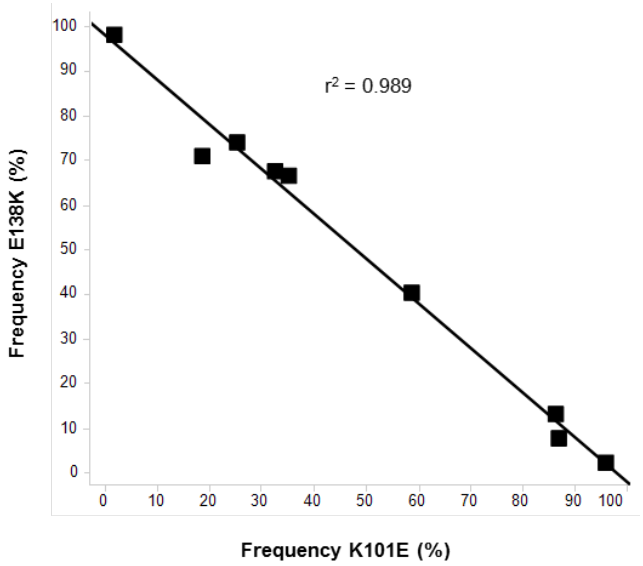


Figure 2.6: Correlation between the observed frequencies of resistance-associated mutations K101E and E138K in HIV RT in samples from 9 HIV-infected subjects. The presence of mutations at these two amino acid positions was derived from paired-end Illumina sequencing.

2.4 Discussion

Recent reviews highlight the growing importance of MPS for different applications in virology [97–99]. A major application is the so-called deep sequencing of targeted regions in order to detect minority variants within the viral population and their possible clinical impact on antiviral therapy [23, 76–79, 100–104]. Until recently, Roche's 454 has been the platform of choice to reliably detect minority variants down to a frequency of 1% [29, 81], which was recently suggested to be sufficient to identify minority variants with clinical impact [82]. However, alternative technologies to address the same application are needed urgently as Roche is planning to fade out the 454 technology by mid-2016. This study evaluated the use of the Illumina sequencing platform for the same application. Using plasmid data it was shown that at least the same or a lower cut-off value of 1% can be applied to Illumina's short-read sequencing technology while maintaining a comparable accuracy. A major difference between the two sequencing platforms is the output volume of sequencing data per run (GAIIx reaches 320 million reads per run while 454 reaches 1 million reads per run). This implies that for a comparable sequencing cost, more than 10-fold higher coverage can be achieved using Illumina's technology, hence suggesting a higher sensitivity compared to 454. Both this study and other comparisons in literature between the two platforms indeed provide experimental support for this statement [90, 105, 106]. Although in theory the higher sequencing depth would enable a lower limit of detection for minority variants as compared to Roche/454, the present study evidences that the cut-off value is mostly bound by error rates introduced during library preparation and sequencing. Recent studies have reported that error correction or data filtering could help reducing the error rate [44, 47, 107–110] suggesting that a lower limit of detection might be feasible by incorporating these methods during variant calling. As analytical sensitivity cannot be infinitely improved by increasing the coverage depth, one could benefit from the large sequencing output volume from the Illumina platform by processing more samples per run at a cost comparable to Roche/454. The experimental set up can be further tailored by running experiments on even higher throughput (HiSeq2000 or HiSeq2500) or lower throughput (MiSeq) Illumina instruments. The reported analysis demonstrate clearly that lowering the coverage depth does not necessarily affect the noise level nor alters the detection accuracy of low frequency variants. A minimal coverage of 2,000x appears to be required, however, in order to ensure an accurate quantitative assessment of minor variant frequencies.

This study demonstrates that the short read technology of Illumina is equally suitable as roche/454 for an in depth characterization of sequence variation at individual codon level. However, the short reads hamper the performance of global haplotype reconstruction as addressed in different studies [90] and limit the direct

detection of linkage between mutations on individual single sequencing reads. The smart use of Illumina's paired-end technology enables linkage analysis beyond the limit of a single sequencing reads. Modulating and selecting the appropriate fragment sizes during library preparation allows detecting linkage in larger fragments – which is only limited by the maximal fragment size that can generate sufficient clustering on the Illumina flow cell, hence similar to or above the sequencing read length generated by 454. A drawback of the approach is, however, the need to define upfront the targeted fragment size, which limits the linkage detection to a priori defined codon positions. In the current study, the existence of linkage between two important IAS-USA resistance-associated mutations (K101E and E138K) in HIV-RT was explored. The paired-end sequencing analysis suggests mutual exclusivity of both mutations, as only 4% of paired sequence reads that cover both codon positions showed linkage between mutations K101E and E138K. Of note, protocols for paired-end sequencing with longer read lengths have meanwhile become available on various Illumina instruments (HiSeq2000: 100 bp, HiSeq2500: 150 bp, MiSeq: 300bp). This implies that contiguous sequences of 550 bp can be achieved on MiSeq by assembling the paired 300 bp reads through a 50 bp overlap, which is a similar size as the single read lengths generated on roche's GS FMX and GS Junior instruments.

2.5 Conclusion

Whereas Roche/454 MPS has for a long time been the method of choice for deep sequencing of viral minority species, in particular because the relatively long read lengths allow identifying linked mutations present on the same viral genome, the Illumina technology was suggested as an interesting alternative for studying low-frequent sequence variants because of its lower cost per sequenced base in combination with the gradually increasing read lengths. The assessment of the performance characteristics of Illumina technology in specific virology applications demonstrates and supports its use as another deep sequencing platform of choice, especially because the higher sequencing output volume from the Illumina platform enables the simultaneous processing of more samples for the same cost as compared to Roche/454 sequencing. However, whereas the high sequencing coverage that can be obtained by Illumina technology results in an accurate estimation of minor variant frequencies, the experimental data show that this does not automatically imply a lower detection limit for the identification of minor variants.

Acknowledgments

The authors thank Carl Van Hove and Elizabeth Van Rossem for the preparation of the samples for sequencing, and Herwig Van Marck and Yves Wetzels for bioinformatics support.

3

Quality Based Adaptive Filtering

B.M.P Verbist, K. Thys, J. Reumers, Y. Wetzels, K. Van Der Borght, W. Talloen, J. Aerssens, L. Clement, O. Thas

VirVarSeq: a low frequency Virus Variant detection pipeline for Illumina Sequencing using adaptive base-calling accuracy filtering. Bioinformatics 2014, doi: 10.1093/bioinformatics/btu587.

Abstract *In virology, massively parallel sequencing (MPS) opens many opportunities for studying viral quasi-species, e.g. in HIV-1 and HCV-infected patients. This is essential for understanding pathways to resistance, which can substantially improve treatment. Although MPS platforms allow in-depth characterization of sequence variation, their measurements still involve substantial technical noise. For Illumina sequencing, single base substitutions are the main error source and impede powerful assessment of low-frequency mutations. Fortunately, base calls are complemented with quality scores that are useful for differentiating errors from the real low-frequency mutations. A variant calling tool, Q-cpileup, is proposed, which exploits the quality scores of nucleotides in a filtering strategy to increase specificity. The tool is imbedded in an open-source pipeline, VirVarSeq, which allows variant calling starting from fastq files. Using both plasmid mixtures and clinical samples we show that Q-cpileup is able to reduce the number of false-positive findings. The filtering strategy is adaptive and provides an optimized threshold for individual samples in each sequencing run. Additionally, linkage information is kept between single-nucleotide polymorphisms as variants are called*

at the codon level. This enables virologists to have an immediate biological interpretation of the reported variants with respect to their antiviral drug responses. A comparison with existing SNP callers reveals that calling variants at the codon level with Q-cpileup results in an outstanding sensitivity while maintaining a good specificity for variants with frequencies down to 0.5%.

3.1 Introduction

RNA viruses such as HIV-1 and HCV exist in their host as complex populations composed of several closely related subgroups. They are referred to as quasispecies and originate from high and error-prone replication rates [10]. This heterogeneous mixture of genomes allows a viral population to rapidly adapt to changing environments. The fittest mutants outcompete the others, allowing the virus to develop resistance to antiviral therapy. The characterization of sequence variation within the viral population is key for understanding pathways to resistance, but the identification of low-frequency variants remains challenging [54, 57].

Until recently, the genetic diversity of a virus population could be assessed only through genotyping by Sanger sequencing, which provides information on only the most abundant viral variants. Massively parallel sequencing technologies allow for a more in-depth characterization of sequence variation, including low-frequency viral strains. However, their measurements still involve substantial technical noise, complicating the analysis [37, 52]. Pyrosequencing, commercialized by Roche 454, was the most common sequencing method for viral population sequencing [53]. The recent announcement by Roche to retract the 454 technology from the market by mid-2016 illustrates the pressing need to evaluate and implement alternative technologies. Recently Illumina's sequencing technique has strengthened its position in this field [67]. Illumina also complements the sequenced nucleotides with quality scores (Q) [56] that reflect the base-calling substitution error probability. The 454 quality scores, however, do not have such an intuitive interpretation [46]. Filtering based on quality scores [64] has already proven valuable to reduce false-positive findings. It often involves the use of a hard quality threshold. Unfortunately, this does not account for variation in quality between runs resulting in too stringent or too relaxed thresholds.

Most variant calling tools focus on the detection of single-nucleotide polymorphisms (SNPs) [44, 47] or perform haplotype reconstruction [45, 66]. Haplotype assembly has its weakness in the detection of low-frequency variants [49], whereas the latter is our main interest. Instead, we prefer to call variants within the read length to avoid challenges encountered in the haplotype reconstruction. On the other hand, linkage between the nucleotides is lost when calling variants at the SNP level. In this contribution we introduce a novel strategy for calling variants at the codon level (nucleotide triplets), which facilitates immediate biological in-

terpretations, particularly in virology applications where drug-target regions are of interest.

In this chapter, we present an innovative approach for variant calling at the codon level, named Q-cpileup, that reduces the number of false-positive findings by exploiting the quality scores of the nucleotides generated by sequencing. Our thresholding strategy is adaptive to provide an optimized threshold for individual samples in each sequencing run. Q-cpileup is imbedded in a pipeline, called VirVarSeq, which starts from fastq files.

3.2 Methods

Several samples were sequenced using Illumina's genome analyzer (GA) IIx according to manufacturing protocols (described in Chapter 2). The VirVarSeq pipeline proceeds as follows:

1. The sequenced reads are aligned against a reference sequence using the Burrows-Wheeler Aligner Tool (BWA) [59].
2. Based on this alignment, a consensus sequence is defined.
3. A realignment is performed against this consensus. This strategy of iterative mapping will increase coverage especially in samples where the consensus strongly deviates from the reference (see appendix A, Figure 1).
4. After alignment, Q-cpileup is executed, which consists of a three-step analysis:
   a) In the first step, the quality scores of the codons in the reading frame of interest are retrieved.
   b) Next the threshold is determined dependent on the quality of the run.
   c) Finally the filtered codon table is constructed.

The VirVarSeq pipeline, which runs from fastq to the filtered codon table, is available at http://sourceforge.net/projects/virtools/?source=directory together with a users guide (see appendix B). All reads containing indel errors are removed before running Q-cpileup. It is hereby assumed that indels will result in non-viable virus. In some rare occasions, however, there might be an insertion mutation at the

Position	Ref Codon	Codon	Count	Coverage	Frequency (%)	Mean Q
001	GGG	AGG	167	18,958	0.88	35
001	GGG	GTG	83	18,958	0.44	16
001	GGG	TGG	15	18,958	0.08	33
001	GGG	GGG	18,693	18,958	98.6	37
002	CGT	CAG	461	19,217	2.4	30
002	CGT	GGG	20	19,217	0.1	5

Table 3.1: Example of a frequency table at the codon level. The different codons observed in a sample are counted at each codon position of the reference genome and their frequencies are calculated using the coverage at that particular position. The quality scores are summarized by averaging the minimum quality scores of the codons. Position: amino acid position of the reference. Ref: codon of the reference genome at a particular position. Codon: codon present in a sample at a particular position. Count: the number of times a particular codon occurs in the cpileup at a particular position. Coverage: the number of reads that fully cover a particular codon position. Frequency: Count/Coverage. Mean: Average of the minimum quality scores of a particular codon at a particular position.

codon level, which can be investigated in a separate analysis (see indel Table appendix A). Below, the different steps from Q-cpileup will be explained in more detail.

3.2.1 Quality of codons

A pileup of read bases is generated using the alignments to a consensus sequence. In analogy with mpileup of samtools, for which the base-pair information at each reference position is described, cpileup describes the codon information at each amino acid position of the reference genome. For each position in the reference genome, the different codons are reported together with one quality score for each codon. This requires that the quality scores of the three nucleotides within a codon have to be summarized. A comparative analysis of different summarizations revealed that the weakest link, i.e. minimum quality score of the three nucleotides in the codon, provided the best separation between low- and high-quality codons (see Appendix A, Figure 2). This minimum quality score represents the codon's nucleotide with the highest probability of being a sequencing error. A codon table is built based on the pileup where for each codon position of the reference the different codons within a sample are reported together with their frequency (Table 3.1). The minimum quality scores of the codons at a particular position are averaged to give a rough idea about the overall quality. However the individual minimum quality scores of the codons themselves is used in subsequent analysis.

3.2.2 Q-intersection threshold (QIT)

The distribution of the minimum quality scores was checked and compared for one particular sample sequenced in two different runs and three different lanes reaching an average coverage of around 30,000 (Figure 3.1). The shape of the distributions can be approximated by a mixture distribution with three truncated normal components (see appendix A for model selection and goodness of fit in appendix A, figure 3). Truncation is performed at the lower and upper ends of the quality score range. The first mode represents a point probability at quality score 2, which is the lowest Illumina quality score. This is due to an artifact created by Illumina's base caller. Read ends with a segment of mostly low quality (Q15 or below) are given a quality score of 2. The second component distribution is a distribution of low quality scores, reflecting the sequencing error distribution. Finally, the highest mode, close to 40, originates from a distribution of reliable calls. Note, that the mixture of three normal components for the errors and reliable calls should be considered only as a working assumption and that neither trimming nor filtering of the data is required before fitting the mixture models. The EM algorithm of McLachlan [60] will be applied for fitting these normal mixture models. We have written an R-wrapper to run the original Fortran code of McLachlan which is embedded in the pipeline VirVarSeq. The EM algorithm was initialized by setting the three modes at 2, 10 and 35 and the variances at 0.8 for the point probability and 40 for the other two distributions. The marginal error probability, the sum of mixing proportion of the distributions at 2 and 10, was set to 15%.

The bulk of quality scores was high, indicating a majority of reliable calls in the dataset (green distribution in Figure 3.1). At the other end, a clear point probability at the quality score of 2 was seen. The red distribution in Figure 3.1 corresponds to low- quality codons that are likely to be sequencing errors. There are several criteria to define a threshold for filtering the low-quality codons and for distinguishing between errors and reliable calls. An approach is chosen that is adaptive and robust. The intersection point of the two component distributions was used and is referred to as the Q-intersection threshold (QIT), which is indicated with vertical dashed lines in Figure 3.1. The distribution of the minimum quality scores of the codons and hence the QIT varies between different runs for the same sample, confirming the need for an adaptive filtering strategy.

3.2.3 Filtering of codon tables

Once the QIT is determined, an updated codon table can be constructed. By default the reads will be trimmed and all codons with a minimum quality score below the threshold will be filtered from the analysis. The influence of trimming is negligible as it mainly affects low-quality nucleotides, which are removed by the filter anyway. The three-step analysis returns a codon table with different variants and their

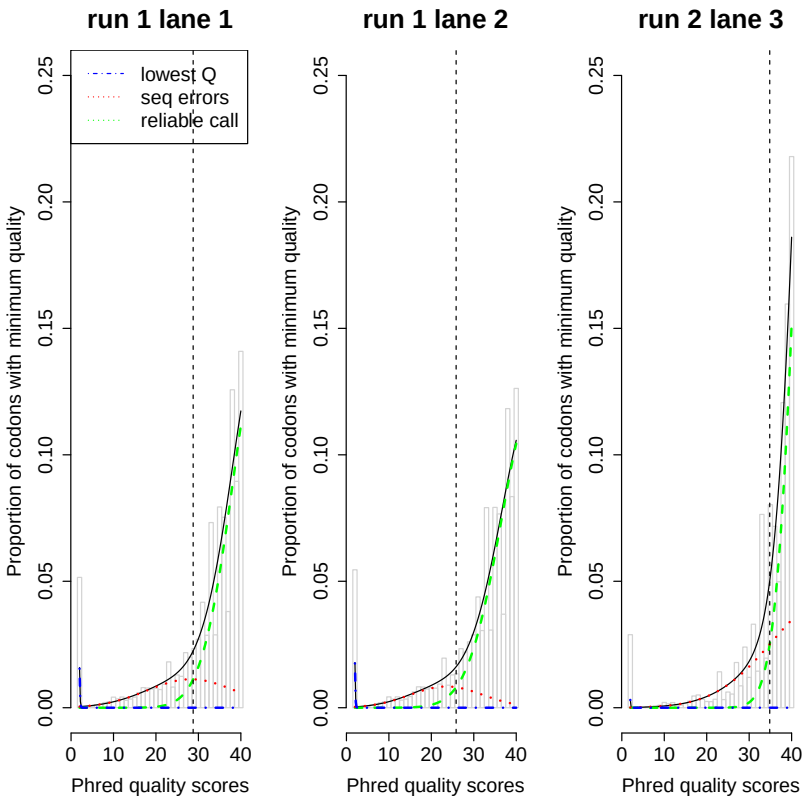


Figure 3.1: Distribution of the minimum quality scores of the codons present in a HCV sample which was sequenced twice in the same run (run1) but in different lanes (lane 1 and 2), and which was sequenced yet another time in another run (run2). The black line shows the overall fit of the mixture distribution, which consists of the blue, the red and the green component distributions. The blue and the red distributions correspond to codons that likely result from sequencing errors, and the green distribution represents reliable calls. The quality intersection threshold (QIT) is indicated with a vertical black dashed line.

frequencies at each codon position of the reference, which is robust to sequencing errors.

3.3 Results

3.3.1 HCV plasmids mixtures

To assess the filtering accuracy of Q-cpileup, we made use of two plasmids that carry HCV NS3 amino acids 1 to 181 and that differ at only two codon positions, 36 and 155. The two plasmids were mixed in four different proportions – 1:10, 1:50, 1:100 and 1:200 – and sequenced with average coverages of 96,211, 81,179, 95,590 and 74,820 respectively. (see Chapter 2 for sample preparation; sequencing data are available at the European Nucleotide Archive, accession number PRJEB5028). A minor variant is defined as a codon that differs from the consensus. Hence, only two minor variants are expected at codon positions 36 [GTC (consensus) $\rightarrow$ ATG] and 155 [CGG (consensus) $\rightarrow$ AAA]; all others can be considered as false-positive findings. For each of the four sequenced mixes, the QIT was determined (Table 3.2). Comparison of the number of variants when no filtering but trimming (QIT=0) is applied and after Q-cpileup reveals that adaptive quality filtering can reduce the number of false-positive findings by 20% to 50% (third and fourth columns of Table 3.2). With Q-cpileup, no false-positive findings are reported with frequencies above 1%, a reporting limit defined in chapter 2. This is in strong contrast with the 7% to 12% FDR without filtering, for discoveries with frequencies above 1%. Q-cpileup is able to reduce the number of false-positive findings while the frequencies of the true minor variants (last columns of Table 3.2) remain unaltered.

The results for the mixing proportion of 1% are shown in more detail in Figure 3.2. A QIT of 19 was used for filtering in this sample (Figure 3.2a). In panel b, the codons equal to the consensus (called major variants) are investigated before (QIT=0) and after filtering (QIT=19). The plasmid data have only two real variants, meaning that the codons investigated here should have frequencies close to 100%. After applying Q-cpileup, the frequencies of the codons are indeed closer to 100, indicating again that the number of false-positive findings is reduced. The minor variants are compared in panel c. Without filtering, several minor variants are reported with frequencies above the reporting limit of 1%. Their frequencies are strongly reduced after filtering, while the estimates of the true minor variant frequencies remain (red triangles). This suggests again that quality score filtering provides effective noise reduction while still retaining the reliable calls at low frequency. Figure 3.2c reveals that our new filtering method allows for lowering the reporting limit, although the discovery of true variants below 0.5% remains challenging.

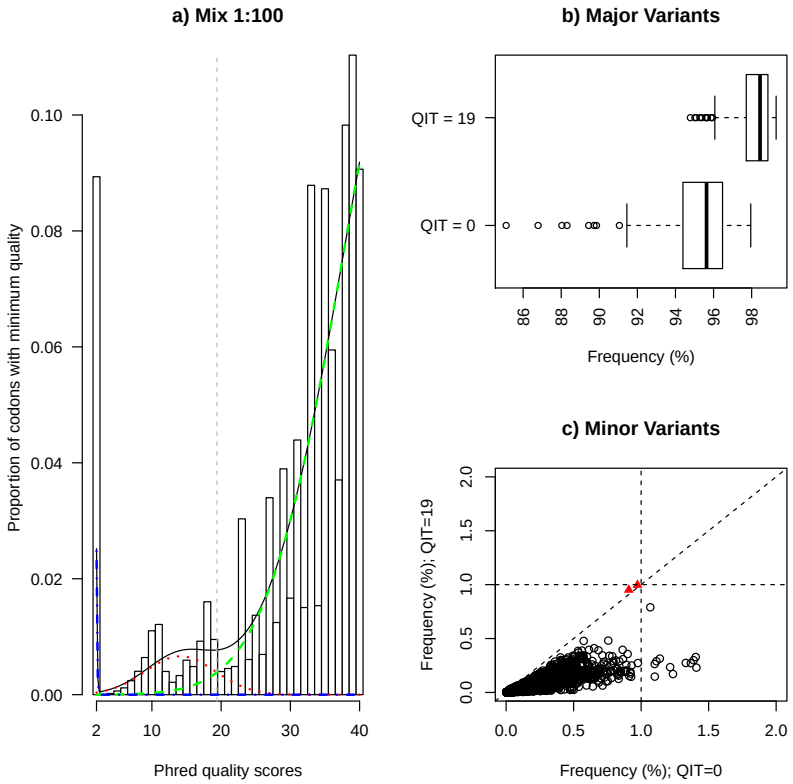


Figure 3.2: a) Distribution of the minimum quality scores of the codons present in the HCV plasmids mixed 1:100. The black line shows the overall fit of the mixture distribution, which consists of the blue, the red and the green component distributions. The blue and the red distributions correspond to codons that likely result from sequencing errors, and the green distribution represents reliable calls. The quality intersection threshold (QIT=19) is indicated with a vertical black dashed line. b) Boxplot of the major variant frequencies when no filtering is applied (QIT=0) and after Q-cpileup with QIT=19. c) Scatterplot of minor variants (frequencies before filtering (QIT=0) on the X-axis and after Q-cpileup on the Y-axis). The true minor variants at codon positions 36 and 155 are indicated with red triangles, all others can be regarded as false-positive findings. The 1% reporting limit is indicated with dotted lines.

Mix	QIT	N° Variants		FDR (%)	
		no filtering	Q-cpileup	no filtering	Q-cpileup
1:10	20	22,405	10,583	12	0
1:50	20	12,724	10,488	7	0
1:100	19	14,886	10,631	6	0
1:200	19	15,692	10,813	8	0

Mix	Freq Codon 36 (%)		Freq Codon 155 (%)	
	no filtering	Q-cpileup	no filtering	Q-cpileup
1:10	10.4	11.6	10.3	11.6
1:50	2.28	2.37	2.25	2.37
1:100	0.97	1.00	0.91	0.95
1:200	0.50	0.49	0.43	0.45

Table 3.2: HCV plasmids results for 181 codon positions sequenced with average depth of 87000. For each of four mixing proportions, the QIT is reported that was used for filtering. The number of reported codons, the false discovery rate for discoveries with frequency above 1% and the estimated frequencies of the true variants are compared before filtering is applied and after Q-cpileup.

3.3.2 Comparison with LoFreq, V-Phaser 2 and ShoRAH

The performance of Q-cpileup is compared with three other methodologies: LoFreq (v0.5.0) [47], V-phaser 2 (v2.0) [44] and ShoRAH (v0.8) [45]. They were run in their default settings and using the previously described plasmid mixtures. With ShoRAH, we were unable to use the original bam file, as unresolvable problems (even after discussion with the developers) were encountered when extracting the reads from the desired region. Therefore, the ShoRAH results are based on a bam file with remapped reads against the reference region of interest. None of the existing methods calls variants immediately at the codon level. Hence, three SNP callers were chosen based on their capabilities for inferring diversity within viral populations. The two codon variant positions present in the mixtures differ at five nucleotide positions, which should be discovered by the SNP callers. The results are presented in the top part of Table 3.3. None of the methods could discover the five SNPs at the 0.5% level, and only LoFreq could retrieve all SNPs present at 1%. Hence, LoFreq is the most sensitive SNP caller in this comparison, but it also detects some false-positive findings with frequency above 1% (bottom rows of Table 3.3). Comparison with Table 3.2 reveals that the sensitivity and specificity for discoveries above 1% with Q-cpileup is outstanding. Especially its sensitivity is of utmost importance: the method is initially developed for finding resistance-associated mutations and missing important variants might mislead further treatment.

	LoFreq			V-Phaser 2			ShoRAH		
SNP (WT)	1:200	1:100	1:50	1:200	1:100	1:50	1:200	1:100	1:50
A (G)	/	1.03	2.41	0.59	1.06	2.37	/	/	2.22
G (C)	0.54	1.01	2.38	/	0.94	2.33	/	/	2.22
A (C)	0.66	1.03	2.16	/	/	/	0.44*	0.80*	1.78*
A (G)	0.48	0.91	2.10	0.52	1.04	2.11	0.44*	0.80*	1.78*
A (G)	/	0.89	2.05	0.48	/	2.07	/	/	1.28
FDR (%)	0.18	0.18	0.18	0	0.18	0	0	0	0
N°	549	553	550	565	578	571	549	546	549

*Table 3.3: Frequency estimates of the 5 SNPs present in the mixture of plasmids with mixing proportions 1:200, 1:100 and 1:50. The first two variants are located in codon 36, while the others from the codon 155. When the actual variant was not discovered it is denoted with /. In case of ShoRAH, the frequency is estimated from three overlapping windows, but often the SNP is only retrieved in two out of the three windows (denoted with *). The bottom rows of the table report the number of codons detected in the NS3 region which in theory should be 548 (543 WT + 5 SNPs). The number of false discoveries with frequencies above 1% is expressed using the false discovery rate calculated as the number of false discoveries with frequencies above 1% divided by the total number of discoveries with frequency above 1%.*

3.3.3 Clinical HCV sample and comparison with 454

Subsequently, Q-cpileup was applied on a clinical HCV sample (see Chapter 2 for sample preparation). The fit of the mixture distribution returned a QIT of 26 (Figure 3.3a). In Figure 3.3b the frequencies of the codons before and after filtering were plotted on the log scale, which allows a better comparison of low-frequency variants. The frequencies of the variants that were removed after filtering were indicated in gray at the bottom. As the truth is unknown in clinical samples one cannot reliably separate true variants from sequencing errors. However, one could compare the discovered variants with 454 sequencing results. As 454 sequencing chemistry is different, another error profile can be expected. Hence, variants that were not discovered with 454 are more likely to be Illumina sequencing errors (and vice versa) and are indicated with red triangles in Figure 3.3b. A good correlation between the filtered and the unfiltered variant frequencies is observed above 1% . The variants that were not detected in the 454 experiment, likely to be false discoveries, drop in frequency after applying Q-cpileup. Hence, our approach seems to control the false discovery rate at a reasonable level up to variant frequencies of 0.5%. A lower coverage depth of the 454 experiment did not allow for comparing Illumina and 454 sequencing for frequencies below 0.1%.

3.3.4 Effect of inter-/intra-run variability on QIT

An equimolar pool of 42 clinical HCV samples was sequenced three times to investigate the variability in sequencing quality and hence the variability of the QIT.

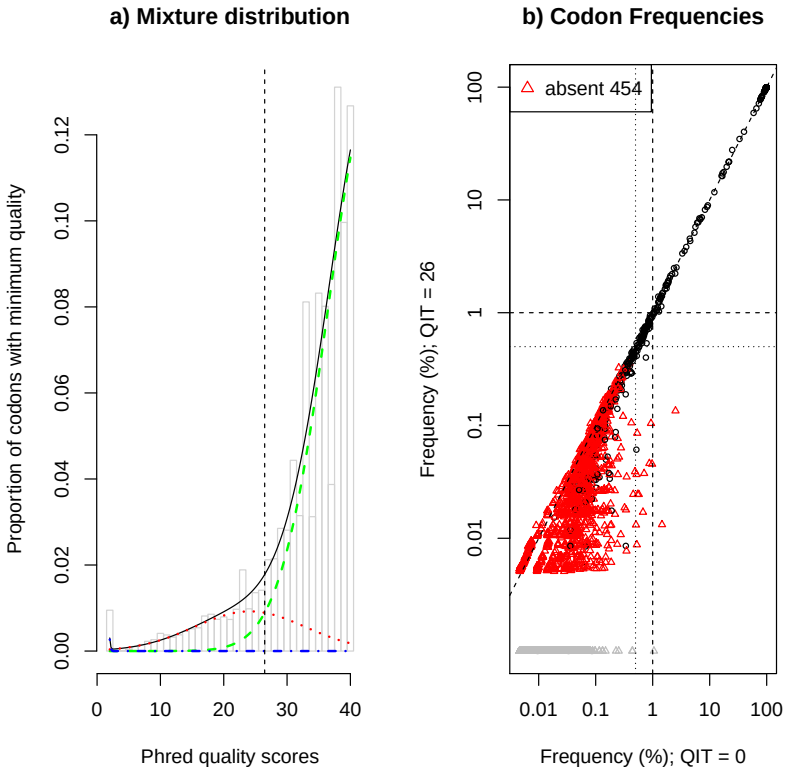


Figure 3.3: a) Distribution of the minimum quality scores of the codons present in a HCV clinical sample. The black line shows the overall fit of the mixture distribution, which consists of the blue, the red and the green component distributions. The blue and the red distributions correspond to codons that likely result from sequencing errors, and the green distribution represents reliable calls. The quality intersection threshold is indicated with a vertical black dashed line. b) Scatterplot of estimated codon frequencies before (QIT=0) and after filtering (QIT=26) on the x-axis and y-axis, respectively. The reporting limit of 1% and 0.5% are indicated with dashed lines and dotted lines, respectively. The codons that were not reported after Q-cpileup are indicated in gray at the bottom. Codons not detected with 454 sequencing were indicated with red triangles and are likely to be false-positive findings.

The 42 samples were sequenced twice within the same run (R1) but on two different lanes (L1 and L2), and they were all sequenced again in another run (R2 L3). For one of these samples, the mixture distribution of each of the three sequencing runs is shown in Figure 3.1. A clear difference in quality between the runs is observed, resulting in different QITs. Boxplots of the QITs of 42 samples for each of the sequencing runs show that the inter-run variability of the QIT is larger than the variability between lanes of the same run (Figure 3.4).

First, the effect of the intra-run variability of the QITs on the final codon table was investigated. The number of reported codons, with frequency above 1% is plotted for both lanes in Figure 3.5a. In all, 74% of the samples report at most one additional codon, depending on the lane on which it is sequenced. The maximum difference in reported number of codons is four. We further explored frequency differences for all codons. The distribution of the differences in frequency between the two lanes is plotted for each sample in Figure 3.5b. Overall, the frequencies are similar with some deviations up to 3%. These maxima, however, are mainly originating from codons located in a GC-rich region, where a coverage drop is observed. The sample where the maximum absolute difference is observed (indicated in red) is investigated in detail in Figure 3.5c. The frequencies for all codons discovered on both lanes are plotted against each other on the log scale. No substantial differences can be observed above 1% . Finally, the maximum differences, for each of the 42 samples, are reported on a relative scale in Figure 3.5d. The sizes of the dots are scaled according to the absolute frequency obtained in lane 1, which teaches us that most deviations occur for frequencies above 50%. Overall, the reported variants and their frequencies are comparable on both lanes after applying Q-cpileup, despite slightly different QITs. This is in strong contrast with the raw data (see Figure 4 in appendix A). These raw data were only trimmed, which is partially based on quality scores. Without Q-cpileup the comparison of the samples sequenced on both lanes reveals that (a) the number of reported variants differs up to 16 variants, (b) deviations of the frequencies go up to 6% and (c) even for the variants with frequency above 1%, some clear deviations between the two lanes can be observed. This suggests that Q-cpileup is able to reduce the number of false-positive findings while retaining the true signal and that the adaptive approach is able to account for differences in quality between the lanes.

In the next step, the inter-run variability of the QITs was investigated. The second run was a very high-quality run with fewer errors (Figure 3.1 comparing run 1 lane 2 and run 2 lane 3). The overall good quality in run 2 makes the estimation of the error component of the mixture challenging and results in a large QIT. The effect of these high QITs on the final codon table was investigated. The number of reported variants, with frequency above 1%, is again similar even after applying these high thresholds (Figure 3.6a). In all, 83% of the samples report at most one additional variant, and only one sample reports four additional variants, which is

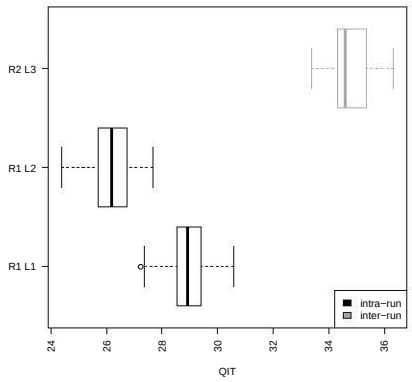


Figure 3.4: Boxplot of the QITs of 42 clinical samples for each of the sequencing runs denoted by run (R) and lane (L) to investigate inter- and intra-run variability. The black boxplots are part of the intra-run comparison, while the gray boxplot represents the QITs of the other run. The inter-run variability is larger compared to the intra-run variability.

again the maximum difference in reported number of variants. The distribution of the frequency differences remains small overall, but the maximum difference in frequency as well as the relative change for these maxima rises (Figure 3.6b and 3.6d). But these maxima occur again at rather high frequencies. When focusing on a particular sample, high QITs seem to have no negative impact on the final codon table (Figure 3.6c). Despite the questionable working assumption that the second component separates the error distribution from the distribution of the reliable calls, the resulting QITs provide reliable codon tables. Hence, our filtering approach seems robust to deviations from the working assumption. Q-cpileup is adaptive and thereby can cope with differences in quality between runs.

3.3.5 Robustness of the method

Approximately 400 samples, from both HCV- and HIV-infected subjects, were sequenced in seven different runs and analyzed with Q-cpileup. Some examples of the HCV results are displayed in Figure 3.3 and 3.5. The sequenced amplicons of HCV samples cover GC-rich regions. It is known that Illumina is error-prone in these regions [37]. This is reflected in the distribution of the minimum quality scores where more low values are observed than with amplicons of HIV samples for which no GC-rich regions were covered (Appendix A, Figure 5). It is especially in these GC-rich regions where you expect that false-positive findings exist with frequencies above 1% as seen in Figure 3.3b on the x-axis. When applying Q-cpileup, the frequencies of these false-positive findings could be reduced. In Figure

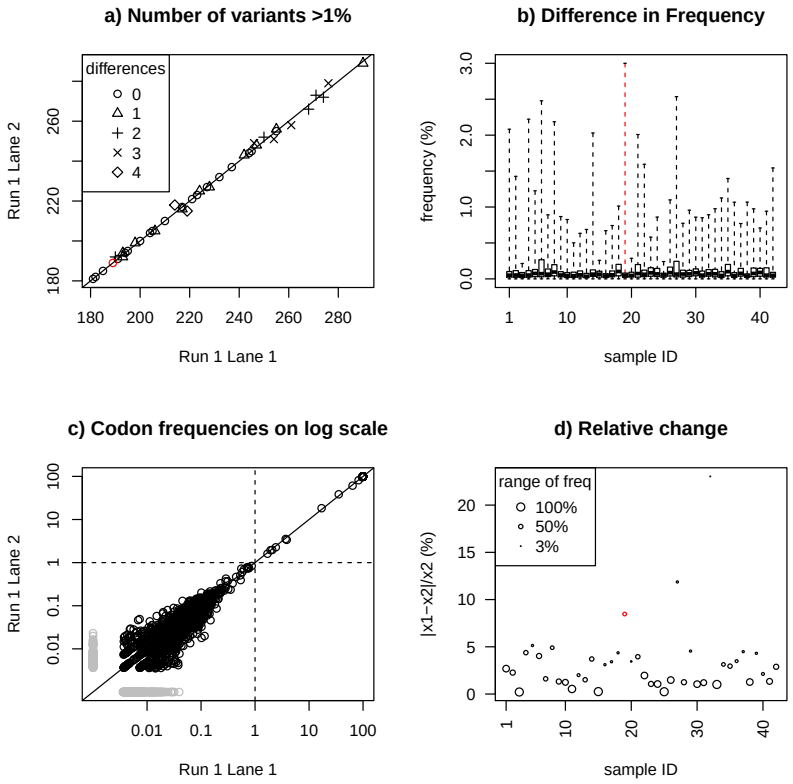


Figure 3.5: Effect of the intra-run variability of the QITs on the final reported codon frequencies. a) Plot of the number of codons with a frequency greater than 1% for the 42 samples sequenced on lane 1 (x-axis) and lane 2 (y-axis). The plotting symbols indicate the number of codons that differ between the two lanes. b) Boxplots of differences in codon frequency between two lanes for each of the 42 samples. For each of the samples, 75% of the codons show differences in frequencies close to zero, while the upper whiskers range roughly between 0.5% and 3% difference in reported codon frequency, depending on the lane where the sample was sequenced. c) Comparison of all codon frequencies on the log scale between two lanes for the sample where the maximum frequency difference is observed. The frequencies of codons not present in the other lane are plotted in gray. d) Relative change for codons with maximum absolute difference plotted for each sample. The relative change is calculated $[x_1\text{-}x_2]/x_2$ with x_1 and x_2 the codon frequencies for lane 1 and lane 2 respectively (Sample with the maximum absolute difference in red). The sizes of the dots are scaled according to the estimated frequency in lane 1, indicating that most maximum differences occur at higher frequencies.

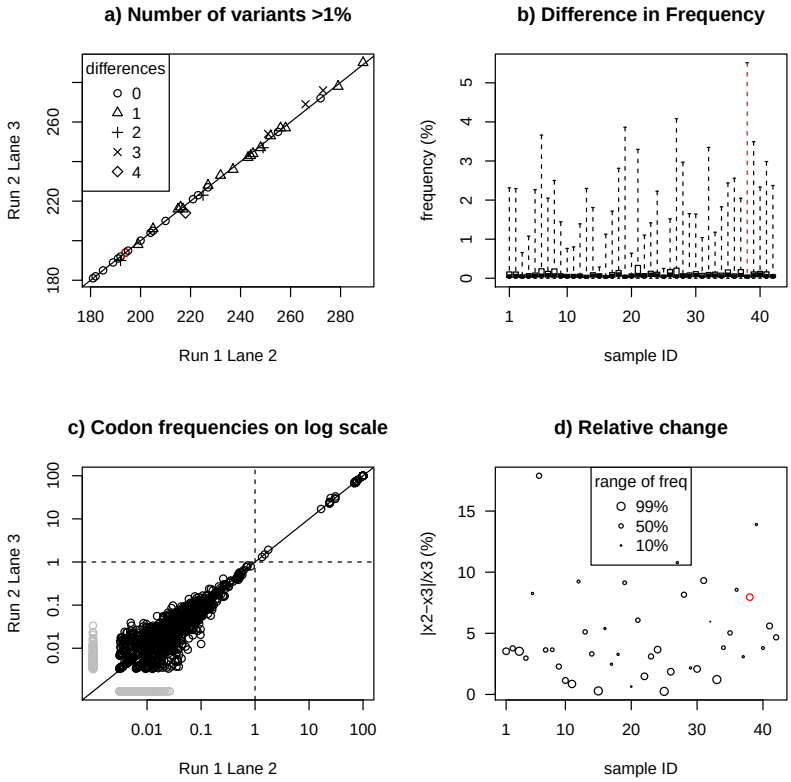


Figure 3.6: Effect of the inter-run variability of the QITs on the final reported codon frequencies. a) Plot of the number of codons with a frequency greater than 1% for the 42 samples sequenced on run 1 lane 2 (x-axis) and on run 2 lane 3 (y-axis). The plotting symbols indicate the number of codons that differ between the two lanes. b) Boxplots of differences in codon frequency between two runs for each of the 42 samples. For each sample, 75% of the codons show differences in frequencies close to zero, while the upper whiskers range roughly between 1% and 5% difference in reported codon frequency, depending on the run where the sample was sequenced. c) Comparison of all codon frequencies on the log scale between the two runs for the sample where the maximum frequency difference is observed. The frequencies of variants that are absent in the other run are plotted in gray. d) Relative change for codons with maximum absolute frequency difference plotted for each sample and scaled according to the estimated frequency in condition 2. The relative change is calculated as $[x_2\text{-}x_3]/x_3$ with x_2 and x_3 the codon frequencies for run 1 lane 2 and run 2 lane 3 respectively (sample with maximum difference is indicated in red).The sizes of the dots are scaled according to the estimated frequency in run 1, indicating that most maximum differences occur at higher frequencies.

5 of appendix A, we illustrate that the proposed strategy also similarly reduces the noise for HIV samples.

Although this new variant calling tool was primarily developed based on Illumina's GAIIx sequencing data, it can also be a valuable tool for other Illumina sequencing platforms such as HiSeq. HiSeq uses the same sequencing-by-synthesis technique and suffers from the same error types [61]. HiSeq data, analyzed with Q-cpileup, are presented in Figure 6 in appendix A. The tool, however, is not applicable for pyrosequencing techniques, such as Roche 454. Their quality scores do not reflect substitution error probabilities but probabilities of calling homopolymers of particular length [46]. For these types of data, the tool can, however, still generate a codon table, but no threshold determination or QIT-based filtering should be applied.

3.4 Discussion

Many sequence variant identification tools have been described in literature to call variants at the SNP level. Most approaches are tailored to call SNPs in human resequencing projects [40] where SNPs can be either heterologous (50%) or homologous (100%). However, in viral deep sequencing projects, the SNP frequency can vary between 100% and 0%, whereas SNPs present in less than 1% of the reads also may be of interest. Further, our main focus is the detection of drug-resistant variants for which a specific drug-target region has to be investigated. Hence, it would be beneficial to report variants at the codon level per amino acid position to enable an immediate biological interpretation of the variants with respect to their antiviral drug responses. Variant calling tools used by the virologists in this research field are either not fully described in-house tools [62] or build based on SNPs [63] at the DNA or RNA level only. In the latter case one needs to retrieve the linkage information between the neighboring nucleotides to deduce effects at the coded amino acid level. This process is not always straightforward. Some of the variant callers developed for viral population sequencing have add-on tools, like V-profiler [58] for V-phaser [44], to convert the list of SNPs to a list of codon variants. However, these are mainly developed for 454 data. None of the available tools reports the variants immediately at the codon level. Therefore, Q-cpileup was initially developed and imbedded in a pipeline. By taking the weakest link as a representative for the quality of the codon, the filtering remains at the individual nucleotide level but reporting is done at the codon level. The approach is adaptive to allow for differences in quality between the runs.

The intersection point between the distributions of the errors and reliable calls is suggested as a threshold (QIT). However, other criteria could also be considered. For instance, the number of false discoveries can be controlled by defining the QIT as a certain quantile of the reliable call component distribution. By using the 5%

quantile as a QIT, 5% of the codons, which are truly present in the population, are falsely considered as being errors. Note that this statement relies heavily on the interpretation of the distributions as errors and reliable calls. However, the three-component mixture distribution should be considered only as a working assumption. Moreover, it is impossible to check the actual interpretation of the mixture distributions. Hence, the interpretation of the different components as error and reliable calls cannot always be warranted, particularly when low quality scores are underrepresented. Therefore, we advise users to assess the distribution plots as diagnostic tools to critically judge whether the chosen threshold is acceptable and/or meaningful.

Instead of applying hard filtering with a QIT, as suggested in this chapter, the posterior probabilities of the reliable call distribution could be used. The counts of a codon at a particular position can be weighted with these probabilities. By doing so, codons with a low quality score will have a low contribution in the final frequency estimates. Hence, data are not filtered but weighted with the probabilities of being truly present in the reliable codon population. This method also relies on the interpretation of the different component distributions as error distribution and reliable call distribution.

Importantly, we have shown that the filtering strategy using hard threshold QIT is robust to runs where the distributions deviate from the working assumption. The impact on the final codon table frequencies was minimal. Overall, our proposed filtering strategy controls the false-positive rate at reasonable levels with no false discoveries above a reporting limit of 1%. The noise is effectively reduced while retaining the reliable calls even at low frequency. This suggests that the reporting limit of detection at 1% could be lowered, although distinguishing true variants from error at 0.5% remains challenging. Depending on the risk one is willing to take to include a small number of false-positive findings, either one of the cutoffs can be used. More importantly, Q-cPileup shows a splendid sensitivity that could not be achieved by the SNP callers while maintaining a good specificity for variants with frequencies down to 0.5%. This sensitivity will allow further investigation of the reported variants above one of the cutoffs defined by the specificity to search for resistance-associated mutations, or in the next step to monitor drug resistance and guide treatment [55]. Currently the clinical cutoff is not yet defined for minority drug-resistant virus variants, and it is still a subject of open debate [65]. Some studies have found no significant association between the presence of low-frequency variants and subsequent virological failure, whereas others report clear correlations [68]. The availability of methods that detect low-frequency variants at the codon level with high sensitivity and good specificity can help in defining the clinical benefit of low-frequency resistance testing.

3.5 Conclusion

A variant calling tool is proposed for identifying true variants at the codon level within a viral population using Illumina sequencing. The variants are filtered using base-calling quality scores for reducing false-positive findings. The lowest quality score of the three nucleotides of the codon is taken as representative for the codon. An adaptive strategy is developed to provide an optimized threshold for individual samples in each sequencing run. The intersection point of the component distributions of the mixture is suggested as a valuable threshold, QIT. Codons with a quality score below this threshold are not reported. The robustness against deviations of the working assumption justifies the utilities of our method for low- and high-quality sequencing runs. It is shown that the generated filtered codon table is reporting far fewer false-positive findings compared with the codon table based on the raw data. Moreover, VirVarSeq has a superb sensitivity compared with existing SNP callers while maintaining a good specificity for codon variants with frequencies down to 0.5%. This suggests that the current reporting limit of detection at 1% can even be lowered. The tool is implemented in a user friendly open-source pipeline, VirVarSeq, which allows virologists to call variants at the codon level starting from the fastq files.

Acknowledgement

The authors wish to thank the scientists at Janssen Pharmaceuticals who collected and produced the data as well as Tobias Verbeke and Joris Meys who gave the necessary IT support. Herwig Van Marck is kindly acknowledged for the insightful discussions. We are grateful to Prof McLachlan and his team for providing the original Fortran code and for giving support to write the R-wrapper. Thanks to Osvaldo, developer of ShoRAH, who was always very helpful when problems were encountered while running ShoRAH.

4

Model Based Clustering

B.M.P Verbist, L. Clement, J. Reumers, K. Thys, A. Vapirev, W. Talloen, Y. Wetzels, J. Meys, J. Aerssens, L. Bijnens, O. Thas

ViVaMBC: estimating Viral sequence Variation in complex populations from Illumina deep-sequencing data using Model-Based Clustering. under revision at BMC bioinformatics (2014) ...

Abstract *Deep-sequencing allows for an in-depth characterization of sequence variation in complex populations. However, technology associated errors may impede a powerful assessment of low-frequency mutations. Fortunately, base calls are complemented with quality scores which are derived from a quadruplet of intensities, one channel for each nucleotide type for Illumina sequencing. The highest intensity of the four channels determines the base that is called. Mismatch bases can often be corrected by the second best base, i.e. the base with the second highest intensity in the quadruplet. A virus variant model-based clustering method, ViVaMBC, is presented that explores the quality scores and second best base calls for identifying and quantifying viral variants. ViVaMBC is optimized to call variants at the codon level (nucleotide triplets) which enables immediate biological interpretation of the variants with respect to their antiviral drug responses. Using mixtures of HCV plasmids we show that our method accurately estimates frequencies down to 0.5%. The estimates are unbiased when average coverages of 25,000 are reached. A comparison with the SNP-callers V-Phaser2, ShoRAH, and LoFreq shows that ViVaMBC has a superb sensitivity and specificity for vari-*

ants with frequencies above 0.4%. Unlike the competitors, ViVaMBC reports a higher number of false-positive findings with frequencies below 0.4% which might partially originate from picking up artificial variants introduced by errors in the sample and library preparation step. ViVaMBC is the first method to call viral variants directly at the codon level. The strength of the approach lies in the modeling of the error probabilities based on the quality scores. Although the use of second best base calls appeared very promising in our data exploration phase, their utility was limited. They provided a slight increase in the sensitivity, which however does not warrant the additional computational cost of running the offline base caller. Apparently a lot of the information is already contained in the quality scores enabling the model based clustering procedure to adjust the majority of the sequencing errors that are correctly called by the second best base. Overall the sensitivity of ViVaMBC is such that the technical constraints like PCR errors start to form the bottleneck for low frequency variant detection.

4.1 Background

In a virology research environment, the study of viral quasispecies in infected patients is essential for understanding pathways to resistance and can substantially improve treatment. Genotypic and phenotypic methods are commonly used for detecting antiviral resistance in clinical HIV-1 and HCV specimens. Standard genotyping such as direct PCR sequencing methods, however, only provides information on the most abundant sequence variants. Modern massively parallel sequencing (MPS) technologies, on the contrary, have the opportunity to allow in-depth characterization of sequence variation in more complex populations, including low-frequency viral strains. However, one of the challenges in the detection of low-frequency viral strains concerns the errors introduced during the sequencing process. As these specific errors may occur at equal or even higher frequencies than true biological mutations, a powerful assessment of low-frequency virus mutations is seriously jeopardized [37, 52].

Many proposals have been made to address this challenge of decreased detection power. Several authors compared the distribution of variants to Poisson, binomial or beta-binomial error distributions [45, 110–113]. They all, however, assume that base calls are of equal quality which is not the case in MPS [37, 114]. As a potential solution other authors suggested to incorporate quality scores when modeling the error distribution [43, 44, 47, 48]. Many of these methods focus primarily on 454 data [43–45, 111–113]. The announcement by Roche to fade out the 454 technology by mid 2016, illustrates the pressing need to focus on alternative technologies [115]. Moreover, the incorporation of quality scores is most appropriate for Illumina sequencing data. Illumina quality scores reflect the base calling substitution error probabilities [56], whereas 454 quality scores do not have

such an intuitive interpretation [46]: they represent the probability of calling a homopolymer up to a particular length.

Illumina's sequencing technology is a sequencing by synthesis technology where the DNA fragments are synthesized one base at a time. The DNA fragments to be sequenced are first spatially separated and amplified, resulting in clusters of identical sequences on the sequencing flow cell. Identification of different bases in the sequencing-by-synthesis process is enabled by using distinct fluorophores for each nucleotide type (A,C,T,G). At every sequencing cycle a single labeled 3'-blocked nucleotide is incorporated to the complementary strand of each DNA fragment. The fluorophore is determined with imaging technology using four different fluorescence channels, one for each nucleotide type. For every fragment in each cycle, the base caller assigns the nucleotide that corresponds with the highest intensity among the four channels. A correct base identification is complicated by multiple effects. On the one hand, emission spectra of the fluorophores are overlapping, especially the A and C intensities and the G and T intensities. On the other hand, phasing and pre-phasing describes the loss of synchrony of the sequence copies of a cluster. Phasing is caused by incomplete removal of the $3'$-protecting groups resulting in sequences within clusters lagging behind in the incorporation cycle. Pre-phasing is caused by the incorporation of nucleotides without effective $3'$-protecting groups. This can cause incorporation of multiple bases in each cycle and might hamper a correct interpretation of the intensities. Quality scores are derived from the intensities [35]. From literature [116] and own experiments it is clear that these quality scores often underestimate the true error probabilities. Extra information which can be used in this context is the second best base calls, which are the bases corresponding to the second highest intensity. Abnizova et al. [50] observed that a mismatch base could often be corrected by its second best base call. In an experiment with known reference sequence 722,505 codons were evaluated of which 34,644 were errors ($\approx$5%). Seventy percent of these errors could be corrected by the second best base calls (see later for more details). Hence, we will explore the utility of second best base calls in addition to the quality scores within a new variant calling algorithm.

Here we propose Virus Variant Model-Based Clustering (ViVaMBC), a method that models error probabilities of the best and second best base calls as a function of the Illumina quality scores. These error probabilities are embedded in a multinomial mixture so that viral variants can be identified and quantified. This chapter will illustrate and validate this method using read sets with known variation and evaluate the minimum sequencing depth. Its performance will be empirically compared with three other methods (LoFreq [47], V-Phaser2 [48] and ShoRAh [45]). Finally, we will demonstrate ViVaMBC on a clinical HCV sample.

4.2 Methods

4.2.1 Experiments

A sample from an HCV-infected patient as well as HCV plasmids were paired-end sequenced using Illumina's genome analyzer(GA)IIx according to manufacturing protocols. A detailed description of the data and protocols is given in chapter 2. The sequencing images are converted into reads using Illumina's off-line base caller (OLB) [117]. In contrast to the standard workflow, using real time analysis (RTA), the OLB can also provide second best base calls which are explored for an improved error correction. In the next steps reads are aligned against a consensus sequence using BWA [59]. The resulting bam files are adapted using GATK clipReads to revert the trimming of the data performed by the aligner, and all reads containing indel errors are removed (workflow presented in Appendix C). It is hereby assumed that indels will result in non-viable viruses. These bam files are used as input of ViVaMBC which is explained in the next section.

4.2.2 Model-based clustering

Let $\mathbf{r}_i$ denote the vector with the best base calls of read i, with i ranging from 1 to n. Similarly, $\mathbf{s}_i$ denotes the vector with the second best base calls of read i. The vector with the corresponding error probabilities is denoted as θ_{ri} and θ_{si} for best and second best base calls respectively. A dummy variable Pair$_i$ is introduced to indicate which end of the DNA segment is sequenced in the paired-end sequencing strategy: $Pair_i$ equals 1 if read i is first in pair and 0 otherwise. In case of single-end experiments the variable $Pair_i$ can simply be omitted from the model. The library of reads represent the whole viral population consisting of several viral subspecies.

The variant calling is applied locally. Upon read alignment the vectors $\mathbf{r}_i$ are retained that cover a small window of the reference sequence under investigation. To avoid the challenges involved in inferring haplotypes [66], only windows smaller than the actual read length are considered. Let m denote the length of the window; thus $\mathbf{r}_i^t = (r_{i1}, \ldots, r_{im})$, AP_i denote the average quality score of read i in window m and θ_{oil} with $l = (1, \ldots, m)$ denotes the probability that the lth nucleotide from read i differs from r_{il} or s_{il}. Typically in our application $m = 3$ to call variants at codon level, nucleotide triplets. ViVaMBC will be a SNP caller when $m = 1$.

Suppose that k variants of length m exist with variant sequences given by the vectors $\mathbf{h}_1, \ldots, \mathbf{h}_k$. Let τ_j denote the prior probability that a read originates from variant j ($j = 1, \ldots, k$). They have the interpretation of relative frequencies of the viral variants within the window, which are the key parameters of interest inferred from the observed data.

The likelihood of the observed data has the natural interpretation of a mixture model with k components that refer to the true variants. The likelihood is the product of the probabilities that a read was generated from the mixture of variants with relative frequencies τ_j:

$$L = \prod_{i=1}^{n} f(\mathbf{r}_i, \mathbf{s}_i) = \prod_{i=1}^{n} \left[\sum_{j=1}^{k} \tau_j f_j(\mathbf{r}_i, \mathbf{s}_i) \right], \tag{4.1}$$

where f denotes a generic density function and f_j is the probability of observing best calls $\mathbf{r}_i$ and second best calls $\mathbf{s}_i$ when read i belongs to variant j. Upon relying on the multinomial distribution, the probability f_j can be written as

$$f_j(\mathbf{r}_i, \mathbf{s}_i) = \prod_{l=1}^{m} f_j(r_{il}, s_{il}) = \prod_{l=1}^{m} \theta_{ril}^{I(r_{il}=h_{jl})} \theta_{sil}^{I(s_{il}=h_{jl})} \theta_{oil}^{(1-I(r_{il}=h_{jl}))(1-I(s_{il}=h_{jl}))}, \tag{4.2}$$

in which $I(A) = 1$ if A is true, and $I(A) = 0$ otherwise. Note, however, that the probabilities θ_{ril}, θ_{sil} and θ_{oil} can not be estimated from the data because the model is over-identified (two parameters for each observation). We therefore model the θ parameters as a function of the quality scores of the best base calls (P_{ril}), a dummy variable Pair ($Pair_i$), and the average quality score (AP_i). For each location l, the θ's refer to a multinomial distribution with three classes for which we suggest a multinomial logit model

$$log \frac{\theta_{cil}}{\theta_{oil}} = \beta_{0c} + \beta_{1c} P_{ril} + \beta_{2c} Pair_i + \beta_{3c} AP_i, \tag{4.3}$$

with $c \in \{r, s\}$. For paired-end experiments eight β parameters need to be estimated together with the variant sequences h_j ($j = 1, \ldots, k$) and the relative frequencies τ_j. For single-end experiments two β parameters are removed since the $Pair_i$ variable can be omitted. To infer the true variants and their frequencies the log likelihood

$$l = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{k} \left(\tau_j \prod_{l=1}^{m} \theta_{ril}^{I(r_{il}=h_{jl})} \theta_{sil}^{I(s_{il}=h_{jl})} \theta_{oil}^{(1-I(r_{il}=h_{jl}))(1-I(s_{il}=h_{jl}))} \right) \right], \tag{4.4}$$

after substituting the θ-parameters with (4.3) will be maximized. However, as closed form solutions for τ_j, h_j and β are not available, numerical methods were implemented for direct maximization of the log likelihood (4.4). The EM algorithm is a popular alternative for maximizing mixture distributions [118, 119]. It requires the introduction of latent or 'missing' indicator variables z_{ij} which are 1 when read i belongs to variant j and zero otherwise. Note that $\mathbf{z}_i = (z_{i1}, \ldots, z_{ik})^t$

are multinomial distributed with density $g(\mathbf{z}_i)$ and $P(z_{ij} = 1) = \tau_j$. Hence the likelihood (4.1) can be augmented

$$L = \prod_{i=1}^{n} f(\mathbf{r}_i, \mathbf{s}_i, \mathbf{z}_i) = \prod_{i=1}^{n} f(\mathbf{r}_i, \mathbf{s}_i|\mathbf{z}_i) g(\mathbf{z}_i) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left(f_j(\mathbf{r}_i, \mathbf{s}_i)\tau_j\right)^{z_{ij}}, \quad (4.5)$$

which in turn allows an efficient factorization by conditioning on $\mathbf{z}_i$. In particular, given $I_{ijl}^{(r)} = I(r_{il} = h_{jl})$ and $I_{ijl}^{(s)} = I(s_{il} = h_{jl})$, the complete data log-likelihood l_c can be written as

$$l_c = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \left\{ \log \tau_j + \sum_{l=1}^{m} \left[I_{ijl}^{(r)} \log \theta_{ril} + I_{ijl}^{(s)} \log \theta_{sil} + (1 - I_{ijl}^{(r)})(1 - I_{ijl}^{(s)}) \log \theta_{oil} \right] \right\}, \quad (4.6)$$

in which the θ parameters have to be substituted with (4.3). The EM algorithm iterates over an expectation (E) and a maximization (M) step until convergence.

1. E step: Computation of the expected complete data log-likelihood (4.6), given the observed data and the current parameter estimates. The solution is given by (4.6) with z_{ij} replaced by

$$\widehat{z}_{ij} = E(z_{ij} \mid \mathbf{r}_i, \mathbf{s}_i) = \frac{\widehat{\tau}_j f_j(\mathbf{r}_i, \mathbf{s}_i|\widehat{h}_j, \widehat{\beta})}{\sum_{l=1}^{k} \widehat{\tau}_l f_l(\mathbf{r}_i, \mathbf{s}_i|\widehat{h}_j, \widehat{\beta})} \quad (4.7)$$

   where f_j depends on $\widehat{h}_j$ and the $\widehat{\beta}$ parameter estimates from the previous M-step.

2. M step: Maximization of the expected complete data log-likelihood from the E-step with respect to τ, $\mathbf{h}$, and β parameters. This results in updated parameter estimates. In particular

   - τ_j is estimated as
   $$\widehat{\tau}_j = \frac{\sum_{i=1}^{n} \widehat{z}_{ij}}{n} \quad (4.8)$$
   - h_j is the most abundant sequence among those with maximal $\widehat{z}_{ij}$ across the variants ($j = 1, \ldots, k$).
   - β parameter estimates are obtained by fitting the multinomial regression model (4.3) using the $\widehat{z}_{ij}$ as weights.

The EM algorithm is initialized with k variants (as a default k is set to 10). The k^{th} most observed variants are taken as initial variant sequences $\mathbf{h}_j$ ($j = 1, \ldots, k$). These variants are updated in each M-step. A variant j will disappear if no sequences are attributed to cluster j. Upon convergence, the number of variants k and their final estimates of τ_j and h_j define the variant population in the window of size m.

The method is optimized for window size $m = 3$ to retain linkage information between single-nucleotide polymorphisms. These nucleotide triplets (codons) facilitate the biological interpretation in the coding regions of the virus. Since resistance-associated mutations against antiviral drugs are particularly of interest, drug-target regions within viral protein coding regions will be investigated. Hence, the reported codon variants can be interpreted immediately with respect to their antiviral drug responses. ViVaMBC is implemented in R and parallelized. Each window of interest can be run on a different core, thereby speeding up the performance. Approximately, one position runs for 1 hour when coverages around 60,000 are reached and $m = 3$. More information can be found in Appendix C.

4.3 Results

In the following sections, first the sensitivity and specificity of ViVaMBC with $m = 3$ will be investigated using read sets with known variation. Subsequently, the minimum depth of coverage needed for unbiased estimates will be defined and its overall performance will be compared with three SNP-callers LoFreq [47], V-Phaser2 [48], and ShoRAH [45]. Finally,ViVaMBC will be illustrated on a clinical HCV sample where the NS3 region will be investigated to search for resistance associated mutations against NS3-4A serine protease inhibitors, telaprevir, and boceprevir [73].

4.3.1 Sensitivity and specificity

Two different plasmids carrying HCV NS3 amino acids 1 to 181 were mixed in four different proportions. These plasmids differ only at codon positions 36 and 155. The mixing proportions were 1:10, 1:50, 1:100, and 1:200 (fastq files are available at the European Nucleotide Archive, accession number PRJEB5028, see chapter 2 for sample preparation). The mixtures were sequenced at an average coverage of 86,000. The plasmid mixture enables the quantification of true positives (variants at the two codon positions) and the assessment of the amount of errors that could be corrected by second best base calls (see Appendix C). The sensitivity of ViVaMBC was quantified using the two variant positions. The estimated frequencies, τ_j, of the real variants at codon positions 36 and 155 were close to the mixing proportions (Table 4.1), suggesting that frequencies down to 0.5% can be

reliably estimated. Codons for the first 181 aminoacids of the NS3 region were called to investigate the specificity. No other variants are expected in this region besides the two variant positions, and hence only the wild type codons (with frequencies close to 100%) and the two variants should be detected. The number of codons reported by ViVaMBC were compared with the number of codons present in the raw data. In analogy with mpileup for SNP calling, a pileup table is built at the codon level where the low-quality parts of the reads are removed prior to the pileup, called trimming (see Appendix C for more details). The comparison with such a pileup table allows to assess the number of false-positive findings that are actually removed by ViVaMBC. The pileup resulted in far more than 10,000 codons while ViVaMBC detected only 599 to 841 codons in the same region (Table 4.2). This indicates that ViVaMBC removes the vast majority of false-positive findings. From the reported codons we removed the wild type codons with frequencies close to 100% together with the two variants and investigated the frequencies of the remaining false-positive findings. The maximum frequency of these errors is above 1% for the pileup and drops below 1% for ViVaMBC. The frequency distribution of the errors is presented in Appendix C Figure 4, which shows that the vast amount of frequencies for false positive variants in ViVaMBC is well below 0.4%. Some false-positive findings are expected in this frequency range as sample and library preparation errors are known to occur with frequencies up to 0.25% [107]. While the discovery of codon variants at 0.5% and 1% was hampered in the pileup table, it could be detected with almost 100% specificity using ViVaMBC. The specific contribution of the second best base error probabilities in ViVaMBC to these increased sensitivity and specificity is further explored in appendix 3.

Mix	36 ATG (%)	155 AAA (%)
1:200	0.45	0.42
1:100	0.92	0.91
1:50	2.28	2.20
1:10	11.04	10.01

Table 4.1: Sensitivity of ViVaMBC in plasmid experiment. Two HCV-plasmids which differ at two codon positions 36 and 155 were combined in a sample for Illumina deep sequencing at four different mixing proportions. Their frequencies were estimated with ViVaMBC, which was able to retrieve codon variants with frequencies up to 0.5%.

4.3.2 Minimum depth of coverage

The influence of coverage depth on the accuracy of $\widehat{\tau_j}$ is investigated using the plasmid data by mixing 1:200 for codon position 155. The original data covered this position 64,668 times. Datasets with lower coverages are generated by random sampling a fraction (f=0.1, 0.2, ..., 0.8 ,0.9) of the reads from the original

Mix	Pileup		ViVaMBC	
	N° Codons	max noise freq (%)	N° Codons	max noise freq (%)
1:200	15,692	1.46	599	0.67
1:100	14,886	1.41	599	0.68
1:50	12,724	1.47	841	0.72
1:10	22,405	1.53	492	0.65

Table 4.2: Specificity of ViVaMBC in plasmid experiment. The number of codons in the NS3 are reported after pileup and ViVaMBC. Theoretically, 183(181+2) codons are expected, but far more are reported, especially when piling up the raw data. The maximum frequency of the false positive codons is presented as well. ViVaMBC is able to reduce these frequencies below 1% while they reached more than 1% after Pileup. This illustrated that ViVaMBC is able to reduce drastically the number of false-positive findings and to lower the detection limit above which 100% specificity is expected.

dataset. Ten datasets were generated for each fraction f resulting in 90 datasets with average coverages ranging between 6,463 and 58,185.

ViVaMBC reported two codons for the original dataset at codon position 155: the wild type codon CGG at a frequency of 99.58% and the variant AAA at 0.42% which is indicated with the green dotted line in Figure 4.1. The frequencies (τ_j) of the variants (h_j) for this position reported by ViVaMBC for each of the 90 re-sampled datasets are plotted in Figure 4.1. The true codon variant AAA (green dots) was detected in all datasets. Averages frequency estimates over the 10 repeats are indicated with green triangles. Figure 4.1 indicates that lower coverages reduce the precision and increase the bias of the estimates. These deviations start to appear from fraction 0.4, which corresponds with coverage around 25,000. The number of false-positive findings also increases when less reads are available, but their frequency estimates remain far below 0.4% and the variant at 0.5% can still be discovered at the lowest coverage.

4.3.3 Comparison with other methods

The performance of ViVaMBC is compared with LoFreq (v0.5.0) [47], V-Phaser 2 (v2.0) [48], and ShoRAH (v0.8) [45] (all ran in their default settings) using the previously described plasmid mixture data. With ShoRAH we were unable to use the original bam file since some problems were encountered when extracting the reads from the desired region. Therefore the ShoRAH results are based on a bam file with remapped reads against the reference region of interest. As none of the existing methods calls variants immediately at the codon level, the evaluation is restricted to the ability to detect variants at individual nucleotide level. The two variant codons differ at 5 nucleotides from the wild type, so 5 SNPs should be detected. The comparison is made with ViVaMBC at the codon level since

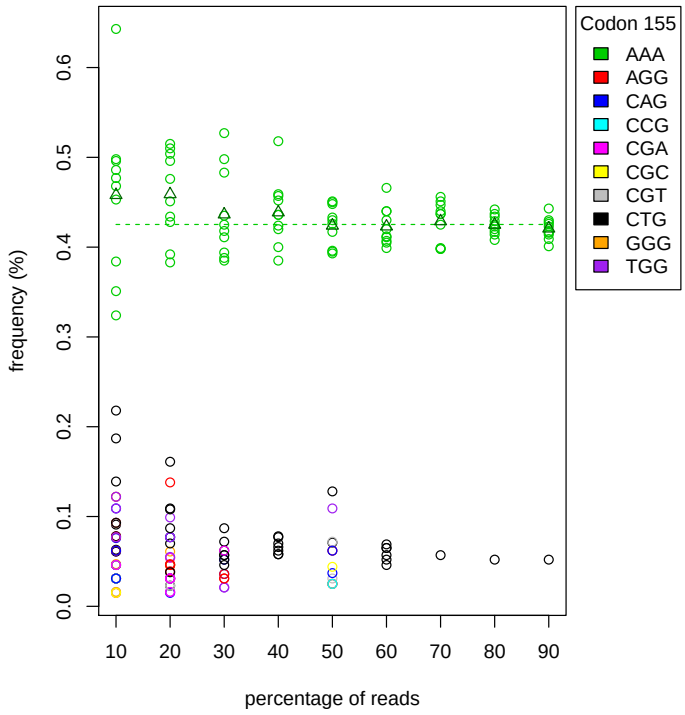


Figure 4.1: Influence of coverage depth on the estimation of τ_j. Datasets with lower coverages are generated by random sampling a fraction (f=0.1, 0.2, . . . , 0.8 ,0.9) of the reads from the original dataset. Ten datasets were generated for each fraction f resulting in 90 datasets with average coverages ranging between 6,463 and 58,185. The reported variants for all re-sampled datasets were plotted and colored according to the discovered codon. The green dots indicate the true variant and all others are false-positive findings. The average frequency of the true variant (averaged over the ten random samples) is indicated with triangles. The dotted line is the true frequency as estimated from the original dataset. Lowering the coverage increases the bias, the variance of the estimate and the number of false-positive findings.

these variants can be interpreted immediately with respect to their antiviral drug responses, which is our primary application domain. The results of ViVaMBC at the SNP level are reported in Appendix C.

The estimated frequencies of the true SNPs for the mixing proportions 1:200, 1:100, and 1:50 are presented in Table 4.3 for the existing methods. None of them were able to retrieve all 5 SNPs at a frequency of 0.5%. LoFreq could recover

them at a frequency of 1% while the others still showed false-negative findings for the 1:100 mixtures. ViVaMBC, on the other hand, was able to discover both codon variants at a frequency of 0.5% and above (Table 4.1).

The total number of false discoveries over the whole NS3 region (181 codons of $3bp$ long) are reported at the bottom of Table 4.3 together with the maximum frequency of these false-positive findings. All methods seem to control the total number of false-positive findings much better than ViVaMBC, but the frequencies of these false-positive findings are close to 1% or even above and hamper the discovery of true variants with similar frequencies. Despite the higher number of false-positive findings discovered in ViVaMBC, a clear distinction between true- and false-positive findings can be made for frequencies around 1%. And with one exception, all false-positive findings fall below 0.4% (see Figure 4.2). So overall ViVaMBC has a higher sensitivity and specificity for the discovery of codon variants at frequencies above 0.5%.

	LoFreq			V-Phaser 2			ShoRAH		
SNP (WT)	1:200	1:100	1:50	1:200	1:100	1:50	1:200	1:100	1:50
A (G)	/	1.03	2.41	0.59	1.06	2.37	/	/	2.22
G (C)	0.54	1.01	2.38	/	0.94	2.33	/	/	2.22
A (C)	0.66	1.03	2.16	/	/	/	0.44*	0.80*	1.78*
A (G)	0.48	0.91	2.10	0.52	1.04	2.11	0.44*	0.80*	1.78*
A (G)	/	0.89	2.05	0.48	/	2.07	/	/	1.28
N° fSNP	3	5	2	19	32	24	4	1	1
Max Freq	1.04	1.01	1.02	0.97	1.40	0.72	0.92*	0.5*	0.89

*Table 4.3: Sensitivity and specificity of competing methods in plasmid experiment. Frequency estimates of the true SNPs after applying the algorithms LoFreq, V-Phaser 2 and ShoRAH on the mixture of plasmids mixed at 1:200, 1:100 and 1:50. Two SNPs should be present in codon 36, while three SNPs are present in codon 155. In case of ShoRAH, the frequency is estimated from three overlapping windows, but often the variant is detected in two out of three windows (denoted with *). None of the methods seem to be able to retrieve all 5 SNPs at 0.5%. The bottom rows of the table report the total number of false SNPs over the whole NS3 region (543 bp long) together with their maximum frequency. The total number of false-positive findings is very low for all methods but their frequencies rise close to 1% which hamper the distinction of true SNPs from this false-positive findings.*

4.3.4 Clinical sample

The application of ViVaMBC is illustrated here on a clinical HCV sample for which the NS3 amino acids 1 to 181 were sequenced with two sequencing platforms (454 and Illumina). The error prone GC-region was used for assessing the performance of ViVaMBC but we compared here the conclusions of the two platforms on the same sample. As 454 sequencing technology uses a different sequencing chemistry (see protocol in chapter 2) it typically results in another error profile. Variants not discovered with 454 can thus be assumed to originate from

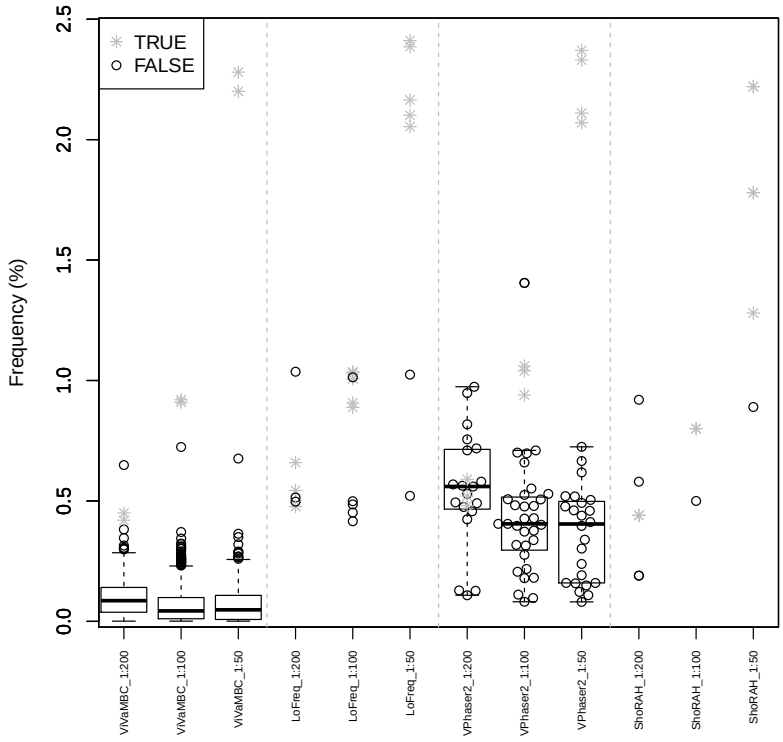


Figure 4.2: Specificity comparison of ViVaMBC with LoFreq, V-phaser 2 and ShoRAH. The frequencies of all minor variants discovered in the three mixtures 1:200, 1:100 and 1:50 are plotted for ViVaMBC, LoFreq, V-phaser 2 and ShoRAH. Note that these variants are at the codon level for ViVaMBC and at the SNP level for the other methods. The false positive variants are indicated with black dots and the true positives with gray crosses. It is clear that although far more false-positive findings are discovered with ViVaMBC, the distinction with the true positives is more apparent.

Illumina sequencing errors (and vice versa). In Figure 4.3a the estimated frequencies of the codons discovered by ViVaMBC are plotted against the corresponding frequencies of the pileup. Codons present in only one of the two methods are plotted in gray on their respective axis. Codons that were not present after piling up the 454 reads were indicated with triangles. Above 0.5% (dotted lines) a good correlation is observed between the two estimates. A few codons with frequencies above 0.5% in the pileup are not reported by ViVaMBC. These codons were also absent in 454 reads and can be considered as false-positive findings in the pileup. On the other hand, three codons showed a frequency above 1% with ViVaMBC while they had a lower frequency in the pileup, one of which was only present in 454. These codons might be false-positive findings called by ViVaMBC, however since it is a clinical sample, it is difficult to assess. Overall, ViVaMBC has a very good sensitivity; none of the true variants discovered with Pileup was missing.

The false discovery rate (FDR), calculated as the number of false-positive findings (codons not present in 454) divided by the total number of discovered codons is investigated for different reporting limits ranging from 0.1% to 1% for both ViVaMBC and pileup (Figure 4.3b). ViVaMBC has much lower FDR compared to pileup table for all reporting limits under investigation. While the FDR rapidly increases at low frequencies for the pileup, it remains stable for ViVaMBC up to a frequency of 0.4% before increasing, which is again in the frequency region where PCR errors start to occur as well. Moreover, the 454 experiment was limited in its detection due to the limited depth of coverage.

Additionally, the three methods LoFreq (v0.5.0) [47], V-Phaser 2 (v2.0) [48], and ShoRAH (v0.8) [45] were ran on the clinical sample. ShoRAH, however, crashed in the final stage of the analysis while running the snv.py script. Hence, Figure 4.4 only presents the comparison of the results of ViVaMBC with those of LoFreq and V-Phaser at SNP level using a barplot representing the number of reported variants at a particular frequency range. The shaded region in the bars for ViVaMBC corresponds to the fraction of codons that were also discovered with 454. Each of the codons reported both by ViVaMBC and 454, contains at least one SNP that should be detected by LoFreq and V-Phaser. V-Phaser, however, reports fewer variants in the majority of the bins, which indicates that it misses some true positives even at higher frequencies. LoFreq seems to perform better and detects all variants up to 1% but is less sensitive at lower frequencies. ViVaMBC probably reports two false positives in the frequency bin $[1\% - 5\%]$, these were also indicated in Figure 4.3a, but our method detects far more true positives especially in low-frequency ranges as compared to the other methodologies. The results confirm that codon variants with frequencies down to 0.5% can be reliably detected with ViVaMBC and that false positives start to appear at lower frequencies. Even down to 0.25% the proportion of false positives remains acceptable.

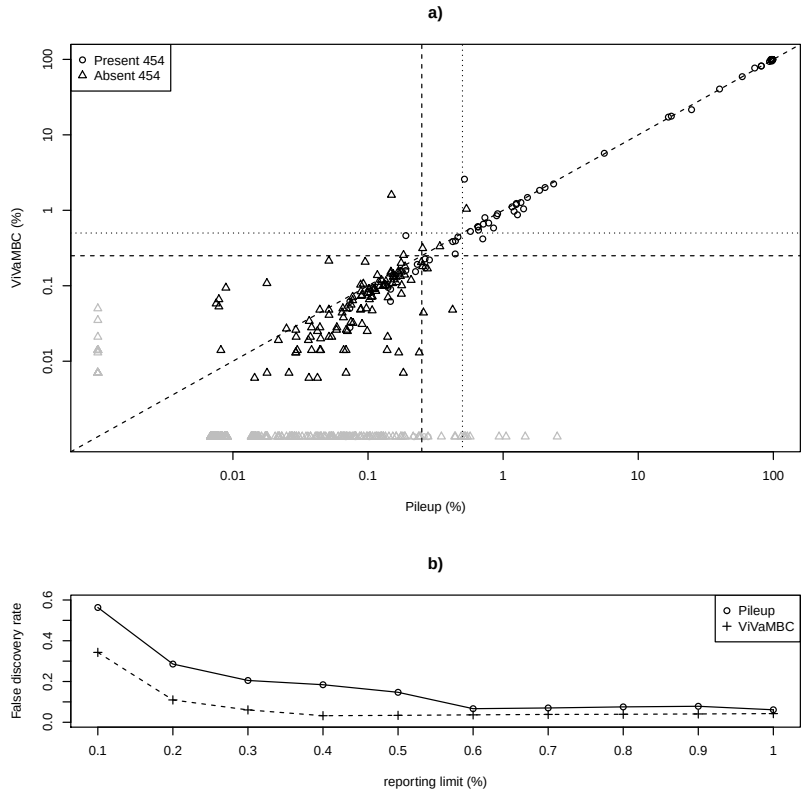


Figure 4.3: Sensitivity and specificity comparison of ViVaMBC with pileup of a clinical HCV sample. a) Comparison of the codon frequencies after piling up the data (x-axis) with the estimated frequencies of ViVaMBC (y-axis). Codons represented with triangles were absent after 454 sequencing on the same sample and hence assumed to be false-positive findings. Codons colored in gray are present in either one of the two methods. Frequencies of 0.5% and 0.25% are indicated with dotted and dashed lines respectively. Above 0.5% and even above 0.25% a good correlation is observed where a few false-positive findings are filtered out using ViVaMBC b) False discovery rates for both ViVaMBC and pileup are calculated with changing reporting limits. The FDR is higher and increases more rapidly for the pileup.

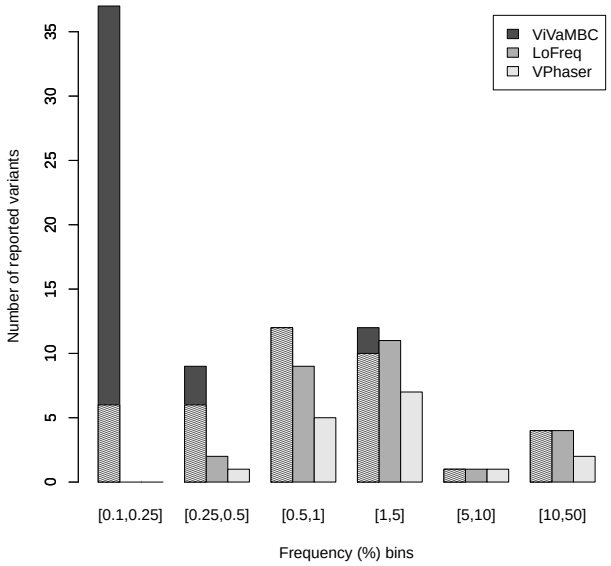


Figure 4.4: Comparison of LoFreq and V-Phaser with ViVaMBC on clinical sample. Barplot represents the number of reported variants (at SNP or codon level) by the different methodologies for different frequency bins. The bars are colored according to the method. The shaded region in the bars for ViVaMBC corresponds to the fraction of codons that were also discovered with 454.

4.4 Discussion

Many SNP calling tools have been described in the literature to correct sequencing errors. Most approaches are however tailored to call SNPs in human resequencing projects [40] where SNPs can only be either heterologous (50%) or homologous (100%). In viral deep sequencing projects, SNPs present in less than 1% of the reads are often of interest [68] making the correction much more challenging. Wilm et al. [47], among others, have shown that incorporating quality scores improves sensitivity without loss of specificity. The comparison with existing tools showed, however, that issues with the detection of low frequency variants with a frequency below 1% in viral populations remains largely unsolved by the currently available methods. ViVaMBC embeds quality scores and the second best base calls within a model-based clustering approach. The method enables an increase in sensitivity for variants with frequencies below 1%, while retaining good specificity above 0.4%. When no second best base calls are available, ViVaMBC still shows an improved sensitivity in comparison with the existing methodologies.

Although the potential of the second best base calls seemed very promising in the data exploration phase, the additional computational cost of running the offline base caller is not warranted for our specific application. At frequencies below 0.4% we start to see errors where some of them are presumed as being incorporated during sample and library preparation. These artificial mutations cannot be identified as errors because the base substitutions are passed to all sequences of the cluster on the flow cell. Hence, these sample and library preparation errors form the limit for detection since only sequencing errors can be corrected with ViVaMBC. To obtain excellent sensitivity and specificity, samples need to be sequenced deep enough. When coverage falls below 25,000 the number of false-positive findings increases and the frequency estimates become biased. Furthermore, ViVaMBC is one of the first tools that calls variants at the codon level, which is particularly of interest in virology applications where drug-target regions are investigated for resistance-associated amino acid mutations. V-phaser 2 and ShoRAH have add-on tools, V-profiler [58] (v1.0) and localVariants [120] (version january 8th 2014), respectively, to convert lists of SNPs to lists of codon variants. LocalVariants is an unpublished tool which is still under development and until now we were unable to run it on our data. At this moment, it failed to define the reading frame based on the number of stop codons. V-Profiler is developed as an add-on tool for V-phaser and the output of V-phaser 2 must be converted to serve as an input for V-profiler. Both V-profiler and localVariants primarily focused on 454 data, and only shifted later to Illumina sequencing. The add-on tools are not fully converted yet, which makes the translation of the list of SNPs to a list of codon variants not straightforward. This illustrates the challenges of retaining linkage information between neighboring SNPs and the need for variant calling methods at the codon level.

The current version of ViVaMBC assumes that each of the n reads covers the entire window of m nucleotides. In practice, many reads cover only partially the window. Although these reads are currently ignored by our method it has a fairly low impact on the results as variant calling is done at the codon level $m = 3$. Ignoring reads can become problematic when larger window sizes m are of interest, for instance when investigating co-occurence of mutations in neighboring codons. If one assumes missingness completely at random, the likelihood approach could be continued with the observed data only. The method only has to be adapted to work with unbalanced data; not all reads will have the same length m. Let v_{il} denote an indicator which is $v_{il} = 1$ if read i has a call at position l and zero otherwise. The density f_j in (4.1) and (4.2) become

$$f_j(\mathbf{r}_i, \mathbf{s}_i) = \prod_{l=1}^{m} f_j(r_{il}, s_{il})^{v_{il}}$$
$$= \prod_{l=1}^{m} \left[\theta_{ril}^{I(r_{il}=h_{jl})} \theta_{sil}^{I(s_{il}=h_{jl})} \theta_{oil}^{(1-I(r_{il}=h_{jl}))(1-I(s_{il}=h_{jl}))}\right]^{v_{il}}. \quad (4.9)$$

Subsequently, the complete data log-likelihood (4.6) becomes

$$l = \log L = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} \left\{ \log \tau_j + \sum_{l=1}^{m} v_{il} \left[I_{ijl}^{(r)} \log \theta_{ril} \right.\right.$$
$$\left.\left. + I_{ijl}^{(s)} \log \theta_{sil} + (1 - I_{ijl}^{(r)})(1 - I_{ijl}^{(s)}) \log \theta_{oil} \right] \right\}. \quad (4.10)$$

We successfully ran ViVaMBC for a HCV-clinical samples where the whole NS3 region is assessed. Investigation of the reported codons will help us to discover mutations associated to resistance against protease inhibitors and to establish the clinical relevance of resistance associated mutations [121]. While ViVaMBC is especially developed for virology applications it might be also applicable in targeted sequencing of cancer associated genes where one wants to uncover the tumor-population heterogeneity. These targeted cancer panels investigate again coding regions, hence working at the codon level makes absolutely sense here.

4.5 Conclusion

ViVaMBC is proposed for identifying variants at the codon level within a viral population using Illumina sequencing. The parameters τ_j and $\mathbf{h}_j$ define the local viral population and are inferred given the observed data. We demonstrated here a superb sensitivity of ViVaMBC while keeping the frequencies of the false-positive findings below 0.4% when an average coverage of 25,000 is reached. The strength of the method lies in modeling the error probabilities, based on the quality scores, which enables to correct a large fraction of the mismatch bases incorporated during the sequencing process. When no second best base calls are available, ViVaMBC can be run without them while it still provides an optimal sensitivity when reporting limits of 0.5% are applied. The technical constraints like PCR errors start to form the bottleneck for low-frequency variant detection.

Acknowledgements

The authors wish to thank the scientists at Janssen Pharmaceutica who collected and produced the data. We are grateful to Tobias Verbeke from OpenAnalytics

who provided the IT infrastructure. Thanks to Osvaldo, developer of ShoRAH, who always was eager to help with some of the issues we encountered when running ShoRAH. This work was supported by Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) who provided a Baekeland mandatory for Bie Verbist [BM100679], Multidisciplinary Research Partnership Bioinformatics: From Nucleotides to Networks Project [01MR0310W] of Ghent University, and IAP research network "StUDyS" grant no. P7/06 of the Belgian government (Belgian Science Policy).

5 Discussion and Perspectives

Two different variant callers were presented in chapters three and four. They approach the challenge of distinguishing noise from low-frequency variants from another perspective. In the following, the two methods will be compared. The presented research project was granted by IWT Baekeland. Hence, the economic finality of the project is important which will be explained in the subsequent section. We close by future perspectives and a final conclusion.

5.1 Qualtity based adaptive filtering versus model based clustering

In contrast to most public available variant callers, the two developed tools, VirVarSeq and ViVaMBC call variants at the codon level which enables immediate biological interpretation of the variants with respect to their antiviral drug responses. They approach, however, the challenge of reducing the false-positive findings during variant calling from two different viewpoints. VirVarSeq is a pipeline which reduces the number of false-positive findings by filtering low-quality codons. A data-driven threshold is defined to account for differences in quality between runs. On contrary, ViVaMBC keeps all data even the low quality codons and models the error probabilities of the best and second best base calls by using the quality scores. These error probabilities are used to divide the codons into clusters. In the following, both methods will be compared on HCV plasmids and a clinical sample.

5.1.1 HCV plasmids mixtures

Both variant calling methods were applied to the mixture of plasmids that carry HCV NS3 amino acids 1 to 181 and that differ only at two codon positions, 36 and 155 (see chapter 3 and 4). These plasmids were mixed in four different proportions –1:10, 1:50, 1:100 and 1:200. The results are brought together and compared in Table 5.1. Both methodologies find the two variants in the NS3 region with frequency estimates close to the mixing proportions suggesting that they have both superb sensitivities. Other methods, as described in chapter 3 and 4 were not able to find the corresponding SNPs at the lowest frequencies.

Mix	VirVarSeq			ViVaMBC	
	QIT	Codon 36	Codon 155	Codon 36	Codon 155
1:10	20	11.6	11.6	11.04	10.1
1:50	20	2.37	2.37	2.28	2.20
1:100	19	1.0	0.95	0.92	0.91
1:200	19	0.49	0.45	0.45	0.42

Table 5.1: Estimated frequencies of the variant positions for the different mixing proportions of the HCV plasmids for both VirVarSeq and ViVaMBC together with the quality threshold used by VirVarSeq.

All variants, except the spiked-in variant at position 36 and 155, can be considered as false-positive findings. After quality-based filtering, more than 10,000 false-positive findings remain while ViVaMBC reduces these number to a few hundreds (displayed on the right of Figure 5.1). This means that VirVarSeq reports almost all possible codons at very low frequencies $(64 * 181 = 11584)$. The distribution of the frequency estimates of false-positive findings is shown in Figure 5.1. The number of outliers for VirVarSeq reaches a thousand, which is more than the total number of false-positive findings for ViVaMBC. No false-positive findings remain with frequencies above 1% for neither one of the methods. False-positive findings start to occur regularly with frequencies around 0.5% for VirVarSeq while ViVaMBC is able to reduce these frequencies a little bit more. Hence, detection of low-frequency variants down to 0.5% becomes feasible for ViVaMBC while keeping very good specificities.

5.1.2 HCV clinical sample

Subsequently the variant calling methods are compared for the genomic region, coding for amino acid 1 to 181 of NS3 of a HCV clinical sample. Prior to filtering of the variants called in the sequencing experiment, VirVarSeq requires the determination of the quality threshold, QIT as described in Chapter 3. Figure 5.2 shows the distribution of the minimum quality scores of the codons present in the sample

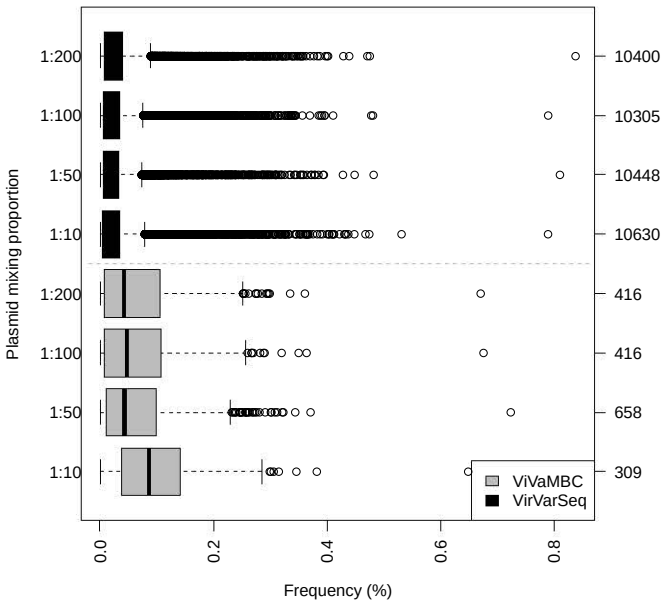


Figure 5.1: Boxplots of false-positive findings for the different datasets both for VirVarSeq (black) and ViVaMBC (grey). Number of false-positive findings represented in the boxplots is displayed at the right.

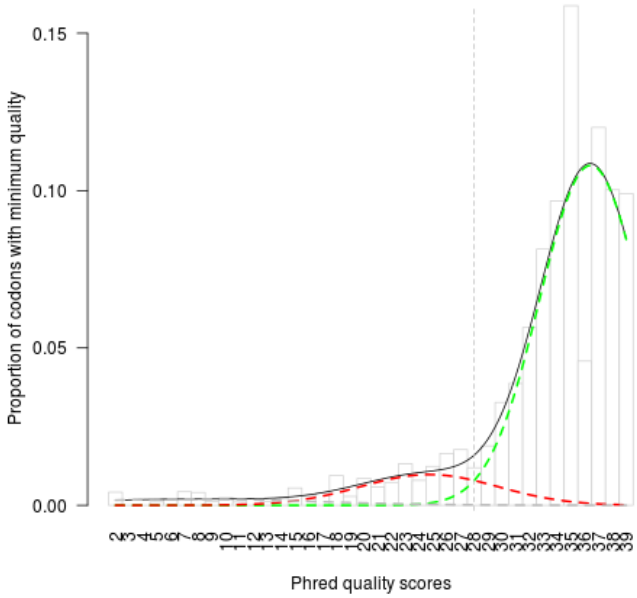


Figure 5.2: Distribution of the minimum quality scores of the codons present in the sample of the whole NS3 region. The black line shows the overall fit of the mixture distribution, which consists of the blue, the red and the green component distributions. The blue (point probability at two), and the red component distribution correspond to codons that likely result from sequencing errors, and the green distribution represents reliable calls. The quality intersection threshold (QIT) is indicated with a vertical black dashed line.

together with the fit of the mixture distribution. The Q-intersection threshold is 28 which is indicated with a vertical dashed line. Hence, all codons with a minimum quality score below 28 are filtered and 1943 codons remain for the whole NS3 region.

The alignment after consensus mapping of the VirVarSeq pipeline was used as a starting point for ViVaMBC. The second best base calls where added and the model based clustering was applied. ViVaMBC reports only 1390 codons for the NS3 region which is a reduction of 28% of the number of reported codons compared to VirVarSeq.

In Figure 5.3 the frequencies of the reported codons are plotted for the model based clustering and the adaptive filtering on the x-axis and the y-axis respectively. The codons that are reported by only one method are plotted in gray on the bottom of the respective axis. It concerns 721 codons for VirVarSeq and 168 codons for ViVaMBC. Above the reporting limit of 1% the two methodologies are in agree-

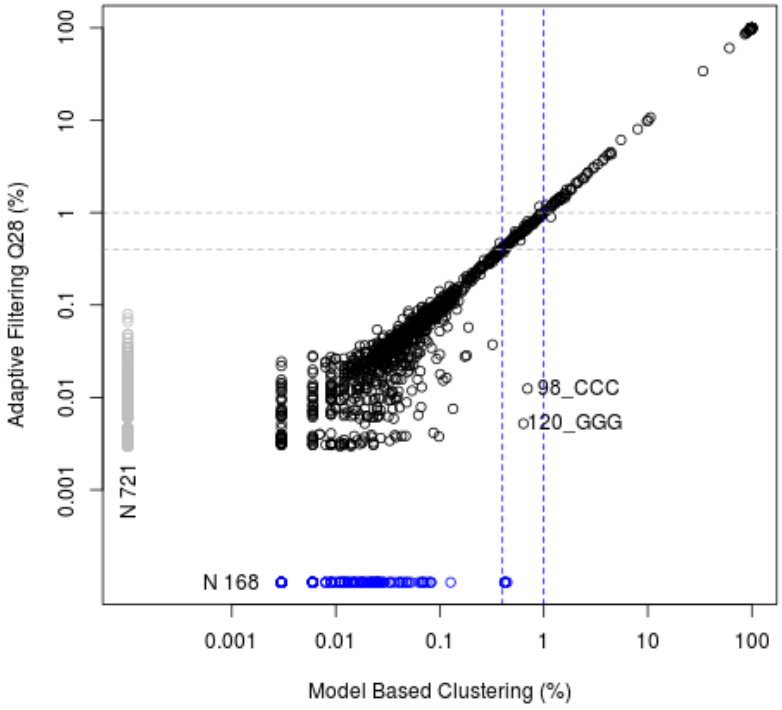


Figure 5.3: Comparison of the codon frequencies reported by ViVaMBC (x-axis) and by VirVarSeq (y-axis). Codons that are only reported by ViVaMBC are plotted along the x-axis in blue, while the ones that are only reported by VirVarSeq are plotted along the y-axis in gray. Reporting limits of 1% and 0.4% are indicated with dashed lines.

ment. Even up to 0.4%, a reporting limit which was suggested for ViVaMBC, the reported codons and their frequencies are similar. Only two codons, both in the GC-rich region which is known to be error-prone, have frequencies that clearly differ between the two methodologies. Additionally six codons have estimated frequencies close to 0.4% in ViVaMBC while they were filtered in VirVarSeq. Since, the methods are applied on a clinical sample it is difficult to judge which of the two methodologies is more correct. Below 0.4%, the frequency estimates of the reported codons differ between the two methodologies where in general the estimates of ViVaMBC are slightly higher.

5.1.3 Discussion

The two methodologies, VirVarSeq and ViVaMBC, report similar codons when the reporting limit of 0.5% is applied, although they approach the challenge of distinguishing noise from low-frequency variants from another perspective. Below the reporting limit, VirVarSeq reports probably more false-positive findings compared to ViVaMBC. In some cases, VirVarSeq reports almost all possible codons. On the other hand, ViVaMBC might introduce false positive findings in the GC-rich region, since frequency estimates of variants in these regions seems to be boosted. Hence, it might be appropriate to add additional parameters in the modeling approach which account for the GC-context. Further it is difficult or either impossible to assess if ViVaMBC, while reporting less ultra low-frequency variants, still detects all reliable variants at these low frequencies. There is most probably a trade-off between sensitivity and specificity where VirVarSeq is most sensitive, but less specific. It is very difficult to decide which of the two elements, sensitivity or specificity, is most important and it might depend on the application. Often, reporting limits are applied above which both splendid sensitivity and specificity are obtained. This makes absolutely sense, since not all variants within the viral population are clinical relevant. The question that rises here again is what is the clinical cut-offs that is needed to define minority drug-resistant variants. Within Janssen Pharmaceuticals the 1% reporting limit was applied, but this can be lowered nowadays to 0.5% when applying VirVarSeq or even down to 0.4% when applying ViVaMBC.

5.2 Valorization

HCV and HIV exist in infected patients as a large viral population of intrahost variants, which may be differentially resistant against antiretroviral drugs. Hence, standard antiretroviral therapy (ART) consists of a combination of drugs to maximally suppress the virus and stop the progression of the disease. The combination of drugs depends on the constitution of the viral population of the infected patients. Therapeutic guidelines therefore recommend genotypic resistance testing before initiating ART [123]. Currently, standard genotyping by Sanger sequencing [22] is used in clinical practice to detect the viral variants. However, Sanger sequencing can only detect viral variants representing more than 15 to 25% of the viral population. A common and clinically relevant question that clinicians ask is: how often this genotyping underestimates the presence of low-frequency variants and whether these low-frequency variants can contribute to treatment failure. Since more sensitive techniques have been developed, including massively parallel sequencing, low-frequency variants can be detected and quantified nowadays down to 0.5 to 1% [23]. However, clinical relevance of detecting low-frequency

variants remain open to debate. According to some studies, low-frequency baseline drug resistance is associated with a higher risk of treatment failure [14–16]. Other studies have not found an influence of minority resistance mutations on the treatment response [21].

One of the reasons why this question remains open to debate might be the challenges involved during the variant calling. The technical noise that is interpreted as a low-frequency variant might disturb the final analysis. Hence, algorithms with high specificity in combination with a high sensitivity can help in defining the clinical relevance of low-frequency variant detection. Existing methodologies, LoFreq [47], V-phaser 2 [44], and ShoRAH [45] showed very high specificities but tend to miss some important discoveries. The presented methodologies VirVarSeq and ViVaMBC showed high sensitivities in combination with good specificities when variants are reported down to 0.5%. Hence, application of these methods can help in determining if a drug-resistant variant at 0.5% is still clinically relevant.

Detection of resistance associated variants against anti(retro)viral drugs represents the economic finality of the project. Since these drugs target certain proteins, consisting of amino acids translated from the viral genome, it is of interest to investigate variants at codon level. Most variant callers described in literature, however, call the variants at the SNP level but here linkage information between the different SNPs is lost. Therefor, we developed variant callers VirVarSeq and ViVaMBC who call variants at the codon level. Hence the economic objective of the project is immediately taken into account.

Currently at Janssen Pharmaceuticals the pipeline VirVarSeq, described in chapter 3, is used to support some of their clinical trials to determine resistance associated mutations. These mutations are put in the label of the drug which guides the clinicians in determining the best combination therapy for each patient. Most samples taken during the clinical trial are sequenced using Sanger. Only those samples from patients that fail therapy and where no known resistance associated mutations were found, are deep-sequenced and analyzed using our pipeline.

5.3 Perspectives

It is clear that massively parallel sequencing opens new routes to study viral diversity and its impact on resistance. A big step forward in the use of these techniques in clinical practice is the recent approval by FDA. End of 2013, FDA granted first marketing authorization for Illumina's MiSeqDX [132]. This recent authorization of a sequencing platform for clinical use will probably expand the use and development of genome-based tests [130, 131] and will promote the further development of personalized medicines. However, MPS techniques and the variant calling tools will not immediately substitute Sanger sequencing in routine clinical practice to

personalize the combination treatment of HIV patients for instance. Before this is going to happen, it is essential to perform large prospective studies to assess the impact of low-frequency variants on virological response; clinical cut-offs need to be defined for minority drug-resistant variants. As said, the developed tools can help here since they have an increased sensitivity versus the existing methodologies. Since Illumina's MiSeq and MiSeqDx is based on the sequencing-by-synthesis approach, similar to Illumina's Genome Analyzer IIx, VirVarSeq and ViVaMBC are expected to work with these technologies as well and hence, their potential can be further exploited.

Illumina is currently the leader in the NGS industry which is of course good news for the application domain of the developed methodologies. However, the rapidly evolving field of MPS [133] is a challenge for the development of such methodologies. To allow application of the developed variant callers to new emerging technologies, the variant callers should be sufficiently generic. Base-calling itself is highly technology dependent, hence our variant callers start only after base-calling and alignment of the reads. As long as the quality scores have an interpretation of substitution error probabilities both methods can be applied and their increased sensitivity versus the existing methodologies can be exploited. The pyrosequencing techniques, like 454 [29] and Ion Torrent [134] are hence not appropriate since the quality scores have an other interpretation. Currently, the single molecule sequencing techniques are emerging. These technologies look promising since they do not need the amplification steps prior to the sequencing and hence an important source of error in the sequencing process can be avoided. The current leader in this field is Pacific Biosiences with its single molecule real time sequencing technique, also known as SMRT [128]. It is, similar to Illuminas technology, a sequencing-by-synthesis technique based on the polymerase chain reaction. Another single molecule sequencing technology is Nanopore sequencing. This technology is based on the transit of a DNA molecule through a pore while the sequence is read out base by base through the effect on an electric current or optical signal [135]. Contradictory, both methods suffer from rather high error rates and further improvement will therefor be required especially when low-frequency variants are of interest. Whether or not these technologies will shake up the sequencing industry or if Illumina will remain the leading technology in the coming years remains to be seen. However, we believe that our variant callers have potential to be adapted to work for these emerging technologies as well.

Additionally, the application domain of VirVarSeq and ViVaMBC is not restricted to virology which increases their potential. The increase in sensitivity to detect low-frequency variants that these methods can offer, will be highly beneficial across many fields. One example is within oncology, where there is a substantial amount of variation within cancer types that affect specific tissues or organs and that leads to different disease outcomes and responses to treatment. Massively

parallel sequencing has emerged here as well as an excellent tool to characterize the individual cancer types in order to better understand the underlying heterogeneity [124, 125]. The somatic point mutations which characterize the cancer genome, occur however at low frequency [126]. Hence highly sensitive variant callers need to be applied [127]. In this sense, a future direction could be to extensively test and fine tune ViVaMBC as a somatic mutation-calling method where coverages are only a hundred up to thousand fold, which is a typical setting for tumor data. These decreased coverages will imply that variants with frequencies lower than 0.5% can not be detected with good specificity. The simulation exercises in chapter 2 and 4, where the frequency estimates are investigated at different coverage, may suggest that variants down to 2% or maybe 1% could be detectable. But further investigation is needed.

5.4 Conclusion

Two variant calling tools, VirVarSeq and ViVaMBC are proposed for identifying variants within viral populations using Illumina sequencing. They both call variants at the codon level to allow for an immediate biological interpretation in terms of drug resistance variants. The way how they overcome the challenge of reducing false-positive findings during variant calling differs and starts from two different viewpoints. VirVarSeq is a quality-based filtering approach where codons below a predefined quality are filtered out. The quality cutoff is defined as the intersection point of two components from a mixture distribution fitted on the quality values of the codons, which is defined as the minimum quality score of the three nucleotides. This data-driven definition of the quality cutoff allows for differences in quality between runs. ViVaMBC, on the other hand, keeps all detected codons but models the error probabilities of the base call and the second best base call by using the quality scores of the individual nucleotides. Based on these error probabilities the codons are divided into clusters where the major variant within the cluster defines the actual variant and where the cluster size is an estimate of the variant frequency.

Comparison of the two methodologies shows that they have both a splendid sensitivity while retaining very good specificities for variant frequencies above 0.5%. Above this limit, both methods report in general the same variants with similar frequencies. However, scientists at Janssen Pharmaceuticals prefer the VirVarSeq pipeline while statisticians are more in favor of ViVaMBC. Filtering approaches are more imbedded in the field of bioinformatics although they are sensitive to the parameter choices. This is partly overcome by applying an adaptive approach where the threshold is defined on each individual data set counting for differences in quality between the runs. However, VirVarSeq still bears the potential risk of biasing the results by removing parts of the data which does not apply for ViVaMBC. Further, the number of reported variants below the report-

ing limits are much lower for ViVaMBC. This would allow to report all variants independent of reporting limits if some false-positive findings were allowed. On the other hand, the performance characteristics of VirVarSeq are much better, both in computing time as well as in development time and hence the more pragmatic approach, VirVarSeq might be preferred especially in an industrial setting.

Bibliography

[1] Dmitri Iwanowski. *Über die Mosaikkrankheit der Tabakspflanze*. Bulletin Scientifique publié par l'Académie Impériale des Sciences de Saint-Pétersbourg/Nouvelle Serie III (St. Petersburg), 35:67–70, 1892.

[2] Robert A Edwards and Forest Rohwer. *Viral metagenomics*. Nature Reviews Microbiology, 3(6):504–510, 2005.

[3] Simon J Anthony, Jonathan H Epstein, Kris A Murray, Isamara Navarrete-Macias, Carlos M Zambrana-Torrelio, Alexander Solovyov, Rafael Ojeda-Flores, Nicole C Arrigo, Ariful Islam, Shahneaz Ali Khan, et al. *A strategy to estimate unknown viral diversity in mammals*. MBio, 4(5):e00598–13, 2013.

[4] Antonio Alcami and Ulrich H Koszinowski. *Viral mechanisms of immune evasion*. Immunology today, 21(9):447–455, 2000.

[5] Welkin E Johnson and Ronald C Desrosiers. *Viral persistence: HIV's strategies of immune system evasion*. Annual review of medicine, 53(1):499–518, 2002.

[6] Nicole Pavio and Michael MC Lai. *The hepatitis C virus persistence: how to evade the immune system?* Journal of biosciences, 28(3):287–304, 2003.

[7] *Fact sheets WHO*. `http://www.who.int/mediacentre/factsheets/fs164/en/`. Accessed: 2014-07-31.

[8] Avidan U Neumann, Nancy P Lam, Harel Dahari, David R Gretch, Thelma E Wiley, Thomas J Layden, and Alan S Perelson. *Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-α therapy*. Science, 282(5386):103–107, 1998.

[9] Xiping Wei, Sajal K Ghosh, Maria E Taylor, Victoria A Johnson, Emilio A Emini, Paul Deutsch, Jeffrey D Lifson, Sebastian Bonhoeffer, Martin A Nowak, Beatrice H Hahn, et al. *Viral dynamics in human immunodeficiency virus type 1 infection*. Nature, 373(6510):117–122, 1995.

[10] Esteban Domingo, Julie Sheldon, and Celia Perales. *Viral quasispecies evolution*. Microbiology and Molecular Biology Reviews, 76(2):159–216, 2012.

[11] J-M Pawlotsky. *Hepatitis C virus population dynamics during infection*. In Quasispecies: Concept and Implications for Virology, pages 261–284. Springer, 2006.

[12] EJJH Domingo and JJ Holland. *RNA virus mutations and fitness for survival*. Annual Reviews in Microbiology, 51(1):151–178, 1997.

[13] Oliver G Pybus and Andrew Rambaut. *Evolutionary analysis of the dynamics of viral infectious disease*. Nature Reviews Genetics, 10(8):540–550, 2009.

[14] Melanie Balduin, Mark Oette, Martin P Däumer, Daniel Hoffmann, Herbert J Pfister, and Rolf Kaiser. *Prevalence of minor variants of HIV strains at reverse transcriptase position 103 in therapy-naive patients and their impact on the virological failure*. Journal of Clinical Virology, 45(1):34–38, 2009.

[15] Karin J Metzner, Stefano G Giulieri, Stefanie A Knoepfel, Pia Rauch, Philippe Burgisser, Sabine Yerly, Huldrych F Gunthard, and Matthias Cavassini. *Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and-adherent patients*. Clinical Infectious Diseases, 48(2):239–247, 2009.

[16] Birgitte B Simen, Jan Fredrik Simons, Katherine Huppler Hullsiek, Richard M Novak, Rodger D MacArthur, John D Baxter, Chunli Huang, Christine Lubeski, Gregory S Turenchalk, Michael S Braverman, et al. *Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment–naive patients significantly impact treatment outcomes*. Journal of Infectious Diseases, 199(5):693–701, 2009.

[17] Vici Varghese, Rajin Shahriar, Soo-Yon Rhee, Tommy Liu, Birgitte B Simen, Michael Egholm, Bozena Hanczaruk, Lisbeth A Blake, Baback Gharizadeh, Farbod Babrzadeh, et al. *Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors.* Journal of acquired immune deficiency syndromes (1999), 52(3):309–315, 2009.

[18] Jurgen Vercauteren, Annemarie MJ Wensing, David AMC van de Vijver, Jan Albert, Claudia Balotta, Osamah Hamouda, Claudia Kücherer, Daniel

Struck, Jean-Claude Schmit, Birgitta Åsjö, et al. *Transmission of drug-resistant HIV-1 is stabilizing in Europe*. Journal of Infectious Diseases, 200(10):1503–1508, 2009.

[19] John M Coffin. *HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy*. Science, 267(5197):483–489, 1995.

[20] Carmen de Mendoza. *HIV drug resistnce testing still important.* AIDS reviews, 16:55, 2014.

[21] Olivia Peuchant, Rodolphe Thiébaut, Sophie Capdepont, Valerie Lavignolle-Aurillac, Didier Neau, Philippe Morlat, François Dabis, Hervé Fleury, Bernard Masquelier, ANRS CO3 aquitaine cohort, et al. *Transmission of HIV-1 minority-resistant variants and response to first-line antiretroviral therapy*. Aids, 22(12):1417–1423, 2008.

[22] Frederick Sanger, Steven Nicklen, and Alan R Coulson. *DNA sequencing with chain-terminating inhibitors*. Proceedings of the National Academy of Sciences, 74(12):5463–5467, 1977.

[23] Chunlin Wang, Yumi Mitsuya, Baback Gharizadeh, Mostafa Ronaghi, and Robert W Shafer. *Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance*. Genome research, 17(8):1195–1201, 2007.

[24] Francis Crick et al. *Central dogma of molecular biology*. Nature, 227(5258):561–563, 1970.

[25] James D Watson, Francis HC Crick, et al. *Molecular structure of nucleic acids*. Nature, 171(4356):737–738, 1953.

[26] Alan D Radford, David Chapman, Linda Dixon, Julian Chantrey, Alistair C Darby, and Neil Hall. *Application of next-generation sequencing technologies in virology*. Journal of General Virology, 93(Pt 9):1853–1868, 2012.

[27] Kary B Mullis and Fred A Faloona. *Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction.* Methods in enzymology, 155:335, 1987.

[28] Phouthone Keohavong and William G Thilly. *Fidelity of DNA polymerases in DNA amplification*. Proceedings of the National Academy of Sciences, 86(23):9253–9257, 1989.

[29] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. *Genome sequencing in microfabricated high-density picolitre reactors*. Nature, 437(7057):376–380, 2005.

[30] Niko Beerenwinkel and Osvaldo Zagordi. *Ultra-deep sequencing for the analysis of viral populations*. Current opinion in virology, 1(5):413–418, 2011.

[31] Dan S Tawfik and Andrew D Griffiths. *Man-made cell-like compartments for molecular evolution*. Nature biotechnology, 16(7):652–656, 1998.

[32] Mostafa Ronaghi, Samer Karamohamed, Bertil Pettersson, Mathias Uhlén, and Pål Nyrén. *Real-time DNA sequencing using detection of pyrophosphate release*. Analytical biochemistry, 242(1):84–89, 1996.

[33] Clyde A Hutchison. *DNA sequencing: bench to bedside and beyond*. Nucleic acids research, 35(18):6227–6237, 2007.

[34] *PhD Thesis Kristof De Beuf*. `http://hdl.handle.net/1854/LU-4161717`. Accessed: 2014-09-15.

[35] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. *Accurate whole human genome sequencing using reversible terminator chemistry*. Nature, 456(7218):53–59, 2008.

[36] *Illumina sequencing-by-synthesis*. `http://openwetware.org/index.php?title=BioMicroCenter:Sequencing\&oldid=739128`. Accessed: 2014-07-31.

[37] Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. *Substantial biases in ultra-short read data sets from high-throughput DNA sequencing*. Nucleic acids research, 36(16):e105–e105, 2008.

[38] *Wikipedia Fastq format*. `http://en.wikipedia.org/wiki/FASTQ_format`. Accessed: 2014-09-07.

[39] Heng Li and Nils Homer. *A survey of sequence alignment algorithms for next-generation sequencing*. Briefings in bioinformatics, 11(5):473–483, 2010.

[40] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. *Genotype and SNP calling from next-generation sequencing data*. Nature Reviews Genetics, 12(6):443–451, 2011.

[41] Daniel C Koboldt, Ken Chen, Todd Wylie, David E Larson, Michael D McLellan, Elaine R Mardis, George M Weinstock, Richard K Wilson, and Li Ding. *VarScan: variant detection in massively parallel sequencing of individual and pooled samples*. Bioinformatics, 25(17):2283–2285, 2009.

[42] Kerensa E McElroy, Fabio Luciani, and Torsten Thomas. *GemSIM: general, error-model based simulator of next-generation sequencing data*. BMC genomics, 13(1):74, 2012.

[43] Christopher Quince, Anders Lanzen, Russell J Davenport, and Peter J Turnbaugh. *Removing noise from pyrosequenced amplicons*. BMC bioinformatics, 12(1):38, 2011.

[44] Alexander R Macalalad, Michael C Zody, Patrick Charlebois, Niall J Lennon, Ruchi M Newman, Christine M Malboeuf, Elizabeth M Ryan, Christian L Boutwell, Karen A Power, Doug E Brackney, et al. *Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data*. PLoS computational biology, 8(3):e1002417, 2012.

[45] Osvaldo Zagordi, Arnab Bhattacharya, Nicholas Eriksson, and Niko Beerenwinkel. *ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data*. BMC bioinformatics, 12(1):119, 2011.

[46] Kristof De Beuf, Joachim De Schrijver, Olivier Thas, Wim Van Criekinge, Rafael A Irizarry, and Lieven Clement. *Improved base-calling and quality scores for 454 sequencing based on a Hurdle Poisson model*. BMC bioinformatics, 13(1):303, 2012.

[47] Andreas Wilm, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. *LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets*. Nucleic acids research, page gks918, 2012.

[48] Xiao Yang, Patrick Charlebois, Alex Macalalad, Matthew R Henn, and Michael C Zody. *V-Phaser 2: variant inference for viral populations*. BMC genomics, 14(1):674, 2013.

[49] M.C.F. Prosperi, Li Yin, David J Nolan, A.D Lowe, M.M. Goodenow, and M.S. Salemi. *Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges*. Scientific reports, 3, 2013.

[50] Irina Abnizova, Steven Leonard, Tom Skelly, Andy Brown, David Jackson, Marina Gourtovaia, Guoying Qi, Rene Te Boekhorst, Nadeem Faruque, Kevin Lewis, et al. *Analysis of context-dependent errors for illumina sequencing*. Journal of bioinformatics and computational biology, 10(02), 2012.

[51] F. Berman et al. *Adaptive Computing on the Grid Using AppLeS*. IEEE Transactions on Parallel and Distributed Systems, 14:369–382, 2003.

[52] Niko Beerenwinkel, Huldrych F Günthard, Volker Roth, and Karin J Metzner. *Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data*. Frontiers in microbiology, 3, 2012.

[53] Niko Beerenwinkel and Osvaldo Zagordi. *Ultra-deep sequencing for the analysis of viral populations*. Current opinion in virology, 1(5):413–418, 2011.

[54] Francisco M Codoñer, Christian Pou, Alexander Thielen, Federico García, Rafael Delgado, David Dalmau, Miguel Álvarez-Tejado, Lidia Ruiz, Bonaventura Clotet, and Roger Paredes. *Added value of deep sequencing relative to population sequencing in heavily pre-treated HIV-1-infected subjects*. PloS one, 6(5):e19461, 2011.

[55] I Dierynck. *Deep Sequencing of the HCV NS3/4A Region Confirms Low Prevalence of Telaprevir-resistant Variants Both at Baseline and End of Study.* Journal of Infectious Disease, accepted, 2014.

[56] Brent Ewing and Phil Green. *Base-calling of automated sequencer traces using phred. II. Error probabilities*. Genome research, 8(3):186–194, 1998.

[57] Sara Gianella and Douglas D Richman. *Minority variants of drug-resistant HIV*. Journal of Infectious Diseases, 202(5):657–666, 2010.

[58] Matthew R Henn, Christian L Boutwell, Patrick Charlebois, Niall J Lennon, Karen A Power, Alexander R Macalalad, Aaron M Berlin, Christine M Malboeuf, Elizabeth M Ryan, Sante Gnerre, et al. *Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection*. PLoS pathogens, 8(3):e1002529, 2012.

[59] Heng Li and Richard Durbin. *Fast and accurate short read alignment with Burrows–Wheeler transform*. Bioinformatics, 25(14):1754–1760, 2009.

[60] GJ McLachlan and PN Jones. *Fitting mixture models to grouped and truncated data via the EM algorithm*. Biometrics, pages 571–578, 1988.

[61] André E Minoche, Juliane C Dohm, Heinz Himmelbauer, et al. *Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems*. Genome Biol, 12(11):R112, 2011.

[62] Marc Noguera-Julian, Maria Casadellà, Christian Pou, Cristina Rodríguez, Susana Pérez-Álvarez, Jordi Puig, Bonaventura Clotet, and Roger Paredes. *Stable HIV-1 integrase diversity during initial HIV-1 RNA Decay suggests*

complete blockade of plasma HIV-1 replication by effective raltegravir-containing salvage therapy. Age, 61(39):49, 2013.

[63] Poornima Parameswaran, Patrick Charlebois, Yolanda Tellez, Andrea Nunez, Elizabeth M Ryan, Christine M Malboeuf, Joshua Z Levin, Niall J Lennon, Angel Balmaseda, Eva Harris, et al. *Genome-wide patterns of intrahuman dengue virus diversity reveal associations with viral phylogenetic clade and interhost diversity*. Journal of virology, 86(16):8546–8558, 2012.

[64] Joke Reumers, Peter De Rijk, Hui Zhao, Anthony Liekens, Dominiek Smeets, John Cleary, Peter Van Loo, Maarten Van Den Bossche, Kirsten Catthoor, Bernard Sabbe, et al. *Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing*. Nature biotechnology, 30(1):61–68, 2012.

[65] M.D Schneider and Sarrazin C. *Antiviral therapy of hepatitis C in 2014: do we need resistance testing?* Antiviral Research, 105:64–71, 2014.

[66] Melanie Schirmer, William T Sloan, and Christopher Quince. *Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes*. Briefings in bioinformatics, page bbs081, 2012.

[67] Kim Thys, Peter Verhasselt, Joke Reumers, Bie M.P Verbist, Bart Maes, and Jeroen Aerssens. *Evaluating the use of the Illumina deep sequencing platform for the detection of minority variants in HIV and HCV.* Journal of Virological Methods, page under review, 2014.

[68] Marie-Anne Vandenhende, Pantxika Bellecave, Patricia Recordon-Pinson, Sandrine Reigadas, Yannick Bidet, Mathias Bruyand, Fabrice Bonnet, Estibaliz Lazaro, Didier Neau, Hervé Fleury, et al. *Prevalence and Evolution of Low Frequency HIV Drug Resistance Mutations Detected by Ultra Deep Sequencing in Patients Experiencing First Line Antiretroviral Therapy Failure*. PloS one, 9(1):e86771, 2014.

[69] Patrick K O'Neil, Guoli Sun, Hong Yu, Yacov Ron, Joseph P Dougherty, and Bradley D Preston. *Mutational analysis of HIV-1 long terminal repeats to explore the relative contribution of reverse transcriptase and RNA polymerase II to viral mutagenesis*. Journal of Biological Chemistry, 277(41):38053–38061, 2002.

[70] Darius Moradpour, François Penin, and Charles M Rice. *Replication of hepatitis C virus*. Nature Reviews Microbiology, 5(6):453–463, 2007.

[71] Michael E Abram, Andrea L Ferris, Wei Shao, W Gregory Alvord, and Stephen H Hughes. *Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication*. Journal of virology, 84(19):9864–9878, 2010.

[72] Victoria A Johnson, Vincent Calvez, Huldrych F Günthard, Roger Paredes, Deenan Pillay, Robert Shafer, Annemarie M Wensing, and Douglas D Richman. *2011 update of the drug resistance mutations in HIV-1.* Topics in antiviral medicine, 19(4):156–164, 2011.

[73] Tarik Asselah and Patrick Marcellin. *New direct-acting antivirals' combination for the treatment of chronic hepatitis C*. Liver International, 31(s1):68–77, 2011.

[74] Evguenia S Svarovskaia, Ross Martin, John G McHutchison, Michael D Miller, and Hongmei Mo. *Abundant drug-resistant NS3 mutants detected by deep sequencing in hepatitis C virus-infected patients undergoing NS3 protease inhibitor monotherapy*. Journal of clinical microbiology, 50(10):3267–3274, 2012.

[75] Sarah Palmer, Mary Kearney, Frank Maldarelli, Elias K Halvas, Christian J Bixby, Holly Bazmi, Diane Rock, Judith Falloon, Richard T Davey, Robin L Dewar, et al. *Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis*. Journal of clinical microbiology, 43(1):406–413, 2005.

[76] Thuy Le, Jennifer Chiarella, Birgitte B Simen, Bozena Hanczaruk, Michael Egholm, Marie L Landry, Kevin Dieckhaus, Marc I Rosen, and Michael J Kozal. *Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use*. PloS one, 4(6):e6079, 2009.

[77] Jeffrey A Johnson, Jin-Fen Li, Xierong Wei, Jonathan Lipscomb, David Irlbeck, Charles Craig, Amanda Smith, Diane E Bennett, Michael Monsour, Paul Sandstrom, et al. *Minority HIV-1 drug resistance mutations are present in antiretroviral treatment–naive populations and associate with reduced treatment efficacy*. PLoS medicine, 5(7):e158, 2008.

[78] BB Simen, MS Baverman, I. Abbate, J. Aerssens, Y. Bidet, O. Bouchez, C. Gabriel, J. Izopet, E. Kessler, F. Di Giallnardo, et al. *An international multicenter study on HIV-1 drug resistance testing by 454 ultra-deep pyrosequencing*. J Virol Methods, 204:31–37, 2014.

[79] Randall Fisher, Gert U van Zyl, Simon AA Travers, Sergei L Kosakovsky Pond, Susan Engelbrech, Ben Murrell, Konrad Scheffler, and Davey Smith. *Deep sequencing reveals minor protease resistance mutations in patients failing a protease inhibitor regimen*. Journal of virology, 86(11):6231–6237, 2012.

[80] Thierry Verbinnen, Herwig Van Marck, Ina Vandenbroucke, Leen Vijgen, Marijke Claes, Tse-I Lin, Kenneth Simmen, Johan Neyts, Gregory Fanning, and Oliver Lenz. *Tracking the evolution of multiple in vitro hepatitis C virus replicon variants under protease inhibitor selection pressure by 454 deep sequencing*. Journal of virology, 84(21):11124–11133, 2010.

[81] André Gilles, Emese Meglécz, Nicolas Pech, Stéphanie Ferreira, Thibaut Malausa, and Jean-François Martin. *Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing*. BMC genomics, 12(1):245, 2011.

[82] Eva Poveda. *Multi-step inhibition against HIV lifecycle-underlying the "magic" of protease inhibitors.* AIDS reviews, 16:52–5, 2014.

[83] Janice Cline, Jeffery C Braman, and Holly H Hogrefe. *PCR fidelity of Pfu DNA polymerase and other thermostable DNA polymerases*. Nucleic Acids Research, 24(18):3546–3551, 1996.

[84] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C Linak, Aki Hirai, Hiroki Takahashi, et al. *Sequence-specific error profile of Illumina sequencers*. Nucleic acids research, page gkr344, 2011.

[85] Susan M Huse, Julie A Huber, Hilary G Morrison, Mitchell L Sogin, D Mark Welch, et al. *Accuracy and quality of massively parallel DNA pyrosequencing*. Genome biol, 8(7):R143, 2007.

[86] Ina Vandenbroucke, Herwig Van Marck, Peter Verhasselt, Kim Thys, Wendy Mostmans, Stéphanie Dumont, Veerle Van Eygen, Katrien Coen, Marianne Tuefferd, and Jeroen Aerssens. *Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications*. Biotechniques, 51(3):167–177, 2011.

[87] Cassandra B Jabara, Corbin D Jones, Jeffrey Roach, Jeffrey A Anderson, and Ronald Swanstrom. *Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID*. Proceedings of the National Academy of Sciences, page 201110064, 2011.

[88] Micah Hamady, Jeffrey J Walker, J Kirk Harris, Nicholas J Gold, and Rob Knight. *Error-correcting barcoded primers allow hundreds of samples to be pyrosequenced in multiplex*. Nature methods, 5(3):235, 2008.

[89] Francisco Rodriguez-Frías, David Tabernero, Josep Quer, Juan I Esteban, Israel Ortega, Esteban Domingo, Maria Cubero, Sílvia Camós, Carles Ferrer-Costa, Alex Sánchez, et al. *Ultra-deep pyrosequencing detects conserved genomic sites and quantifies linkage of drug-resistant amino acid changes in the hepatitis B virus genome*. PloS one, 7(5):e37874, 2012.

[90] Osvaldo Zagordi, Martin Däumer, Christian Beisel, and Niko Beerenwinkel. *Read length versus depth of coverage for viral quasispecies reconstruction*. PloS one, 7(10):e47046, 2012.

[91] V Lohmann, F Körner, J-O Koch, U Herian, L Theilmann, and R Bartenschlager. *Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line*. Science, 285(5424):110–113, 1999.

[92] Nicole Krieger, Volker Lohmann, and Ralf Bartenschlager. *Enhancement of hepatitis C virus RNA replication by cell culture-adaptive mutations*. Journal of virology, 75(10):4614–4624, 2001.

[93] S Dumont, B Fevery, K Van den Brande, V Baumer, H Ceulemans, H DeWolf, LJ Stuyver, and D Koletzki. *Development of a platform to detect drug resistance mutations in the non-structural protein region of the hepatitis C virus Genotypes 1,2,3 and 4.* Journal of Hepatology, 50:S124, 2009.

[94] Hans De Wolf, Herwig Van Marck, Wendy Mostmans, Kim Thys, Ina Vandenbroucke, Veerle Van Eygen, Theresa Pattery, Peter Verhasselt, and Jeroen Aerssens. *HIV-1 nucleotide mixture detection in the virco< sup>®</sup> TYPE HIV-1 genotyping assay: A comparison between Sanger sequencing and 454 pyrosequencing*. Journal of virological methods, 175(1):129–132, 2011.

[95] Phuong Nguyen, Jing Ma, Deqing Pei, Caroline Obert, Cheng Cheng, and Terrence L Geiger. *Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire*. BMC genomics, 12(1):106, 2011.

[96] Mingkun Li and Mark Stoneking. *A new approach for detecting low-level mutations in next-generation sequence data*. Genome Biol, 13(5):R34, 2012.

[97] MR Capobianchi, E Giombini, and G Rozera. *Next-generation sequencing technology in clinical virology*. Clinical Microbiology and Infection, 19(1):15–22, 2013.

[98] Alan D Radford, David Chapman, Linda Dixon, Julian Chantrey, Alistair C Darby, and Neil Hall. *Application of next-generation sequencing technologies in virology*. Journal of General Virology, 93(Pt 9):1853–1868, 2012.

[99] Luisa Barzon, Enrico Lavezzo, Valentina Militello, Stefano Toppo, and Giorgio Palù. *Applications of next-generation sequencing technologies to diagnostic virology*. International journal of molecular sciences, 12(11):7861–7884, 2011.

[100] Gabriella Rozera, Isabella Abbate, Alessandro Bruselles, Crhysoula Vlassi, Gianpiero D'Offizi, Pasquale Narciso, Giovanni Chillemi, Mattia Prosperi, Giuseppe Ippolito, and Maria R Capobianchi. *Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations*. Retrovirology, 6(1):15, 2009.

[101] S Gianella, W Delport, ME Pacold, A Jason, JY Choi, SJ Little, D Douglas, SLK Pond, DM Smith, JA Young, et al. *Detection of minority resistance during early HIV-1 infection: natural variation and spurious detection rather than transmission and evolution of multiple viral variants*. J Virol, 85:8359–8367, 2011.

[102] Charlotte Hedskog, Mattias Mild, Johanna Jernberg, Ellen Sherwood, Göran Bratt, Thomas Leitner, Joakim Lundeberg, Björn Andersson, and Jan Albert. *Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing*. PloS one, 5(7):e11345, 2010.

[103] Peter Messiaen, Chris Verhofstede, Ina Vandenbroucke, Sylvie Dinakis, Veerle Van Eygen, Kim Thys, Bart Winters, Jeroen Aerssens, Dirk Vogelaers, Lieven J Stuyver, et al. *Ultra-deep sequencing of HIV-1 reverse transcriptase before start of an NNRTI-based regimen in treatment-naive patients*. Virology, 426(1):7–11, 2012.

[104] Michael Lauck, Mónica V Alvarado-Mora, Ericka A Becker, Dipankar Bhattacharya, Rob Striker, Austin L Hughes, Flair J Carrilho, David H O'Connor, and João R Rebello Pinho. *Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep pyrosequencing*. Journal of virology, 86(7):3952–3960, 2012.

[105] Jonathan Z Li, Brad Chapman, Patrick Charlebois, Oliver Hofmann, Brian Weiner, Alyssa J Porter, Reshmi Samuel, Saran Vardhanabhuti, Lu Zheng, Joseph Eron, et al. *Comparison of Illumina and 454 Deep Sequencing in Participants Failing Raltegravir-Based Antiretroviral Therapy*. PloS one, 9(3):e90485, 2014.

[106] Justine Cheval, Virginie Sauvage, Lionel Frangeul, Laurent Dacheux, Ghislaine Guigon, Nicolas Dumey, Kevin Pariente, Claudine Rousseaux, Fabien Dorange, Nicolas Berthet, et al. *Evaluation of high-throughput sequencing for identifying known and unknown viruses in biological samples*. Journal of clinical microbiology, 49(9):3268–3275, 2011.

[107] Osvaldo Zagordi, Rolf Klein, Martin Däumer, and Niko Beerenwinkel. *Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies*. Nucleic acids research, 38(21):7400–7409, 2010.

[108] Xiao Yang, Sriram P Chockalingam, and Srinivas Aluru. *A survey of error-correction methods for next-generation sequencing*. Briefings in bioinformatics, 14(1):56–66, 2013.

[109] Zhi Wei, Wei Wang, Pingzhao Hu, Gholson J Lyon, and Hakon Hakonarson. *SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data*. Nucleic acids research, 39(19):e132–e132, 2011.

[110] Patrick Flaherty, Georges Natsoulis, Omkar Muralidharan, Mark Winters, Jason Buenrostro, John Bell, Sheldon Brown, Mark Holodniy, Nancy Zhang, and Hanlee P Ji. *Ultrasensitive detection of rare mutations using next-generation targeted resequencing*. Nucleic acids research, page gkr861, 2011.

[111] Nicholas Eriksson, Lior Pachter, Yumi Mitsuya, Soo-Yon Rhee, Chunlin Wang, Baback Gharizadeh, Mostafa Ronaghi, Robert W Shafer, and Niko Beerenwinkel. *Viral population estimation using pyrosequencing*. PLoS computational biology, 4(5):e1000074, 2008.

[112] Osvaldo Zagordi, Lukas Geyrhofer, Volker Roth, and Niko Beerenwinkel. *Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction*. Journal of computational biology, 17(3):417–428, 2010.

[113] Mattia CF Prosperi and Marco Salemi. *QuRe: software for viral quasispecies reconstruction from next-generation sequencing data*. Bioinformatics, 28(1):132–133, 2012.

[114] William Brockman, Pablo Alvarez, Sarah Young, Manuel Garber, Georgia Giannoukos, William L Lee, Carsten Russ, Eric S Lander, Chad Nusbaum, and David B Jaffe. *Quality scores and SNP detection in sequencing-by-synthesis systems*. Genome research, 18(5):763–770, 2008.

[115] *Announcement454*. `http://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business`. Accessed: 2014-07-31.

[116] Héctor Corrada Bravo and Rafael A Irizarry. *Model-Based Quality Assessment and Base-Calling for Second-Generation Sequencing Data*. Biometrics, 66(3):665–674, 2010.

[117] *Manual Illumina*. http://supportres.illumina.com/documents/myillumina/ec3129a6-b41f-4d98-963f-668391997f1a/olb_194_userguide_15009920d.pdf. Accessed: 2014-07-31.

[118] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[119] Chris Fraley and Adrian E Raftery. *Model-based clustering, discriminant analysis, and density estimation*. Journal of the American Statistical Association, 97(458):611–631, 2002.

[120] *LocalVariants*. https://github.com/ozagordi/LocalVariants. Accessed: 2014-07-31.

[121] Philippe Halfon and Stephen Locarnini. *Hepatitis C virus resistance to protease inhibitors*. Journal of hepatology, 55(1):192–206, 2011.

[122] V Choulakian, RA Lockhart, and MA Stephens. *Cramér-von Mises statistics for discrete distributions*. Canadian Journal of Statistics, 22(1):125–137, 1994.

[123] Melanie A Thompson, Judith A Aberg, Jennifer F Hoy, Amalio Telenti, Constance Benson, Pedro Cahn, Joseph J Eron, Huldrych F Günthard, Scott M Hammer, Peter Reiss, et al. *Antiretroviral treatment of adult HIV infection: 2012 recommendations of the International Antiviral Society–USA panel*. Jama, 308(4):387–402, 2012.

[124] Magdalena Skipper. *Cancer genomics: Indicators for drug response from sequencing*. Nature Reviews Genetics, 13(8):520–520, 2012.

[125] Maureen Cronin and Jeffrey S Ross. *Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology*. Biomarkers in medicine, 5(3):293–305, 2011.

[126] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, et al. *Absolute quantification of somatic DNA alterations in human cancer*. Nature biotechnology, 30(5):413–421, 2012.

[127] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples*. Nature biotechnology, 31(3):213–219, 2013.

[128] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. *Real-time DNA sequencing from single polymerase molecules*. Science, 323(5910):133–138, 2009.

[129] Miles P Davenport, Liyen Loh, Janka Petravic, and Stephen J Kent. *Rates of HIV immune escape and reversion: implications for vaccination*. Trends in microbiology, 16(12):561–566, 2008.

[130] Saumya Pant, Russell Weiner, and Matthew J Marton. *Navigating the Rapids: The Development of Regulated Next-Generation Sequencing-Based Clinical Trial Assays and Companion Diagnostics*. Frontiers in oncology, 4, 2014.

[131] Karen Bijwaard, Jennifer S Dickey, Kellie Kelm, and Zivana Tezak. *The first FDA marketing authorizations of next-generation sequencing technology and tests: challenges, solutions and impact for future assays*. Expert review of molecular diagnostics, (0):1–8, 2014.

[132] Francis S Collins and Margaret A Hamburg. *First FDA authorization for next-generation sequencer*. New England Journal of Medicine, 369(25):2369–2371, 2013.

[133] Erwin L van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. *Ten years of next-generation sequencing technology*. Trends in genetics, 30(9):418–426, 2014.

[134] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards, et al. *An integrated semiconductor device enabling non-optical genome sequencing*. Nature, 475(7356):348–352, 2011.

[135] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. *Continuous base identification for single-molecule nanopore DNA sequencing*. Nature nanotechnology, 4(4):265–270, 2009.

[136] Isabelle Malet, Magali Belnard, Henri Agut, and Annie Cahour. *From RNA to quasispecies: a DNA polymerase with proofreading activity is highly recommended for accurate assessment of viral diversity*. Journal of virological methods, 109(2):161–170, 2003.

A

Supplementary Material. Quality Based Adaptive Filtering.

A.1 Sample preparation

The preparation of the HCV-NS3 plasmids as well as the HCV clinical samples can be found in the material and methods section of chapter 2. The sequencing protocols for Illumina and 454 are also described in this chapter.

A.2 Indel table

All reads containing indels are investigated in a separate analysis dedicated to discover rare codon insertions. The insertions and deletions that occur within these reads are listed in a separate pileup table, referred to as indel table. It reports the actual insertion or deletion and the number of times it was observed for each position where an indel occurred. The frequency of the insertion is calculated using the coverage at the previous position, the frequency of the deletion is calculated using the coverage of the actual position. No filtering based on quality values is applied on the indel table. Composing the indel table is performed in the codon_table.pl script and the results are written in the sub-directory results/codon_table as a separate file.

A.3 Model selection

Mixture distributions with different number of components were fitted on the minimum quality scores of the codons for 5 HCV samples, sequenced on 5 different sequencing runs (Figure A.3). Each HCV sample is colored differently. The goodness of fit (GoF) was assessed using the Cramér-von Mises statistic [122] where the probabilities from the fitted mixture distribution are compared to the empirical probabilities; the smaller this statistic, the better the fit. Mixture distributions with 3 and 4 components show a better fit compared with a lower or higher number of components (Figure A.3a). Two hundred HCV samples were investigated to decide between fitting a mixture distribution with 3 or 4 components (Figure A.3a). The goodness of fit of these 2 models is comparable. Hence, we have chosen to fit the more parsimonious mixture model with 3 components. Moreover, the 3 component model has a more straightforward biological interpretation: a point probability around 2 to account for the data manipulation performed by Illumina, an error distribution and a distribution of the reliable calls.

A.4 Supplementary figures

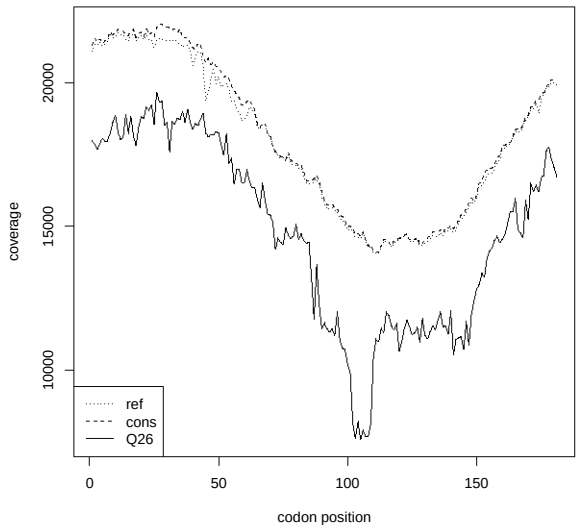


Figure A.1: Coverageplot for NS3 region of a clinical HCV sample where the consensus clearly deviates from the wild type. The coverage is plotted with dotted, dashed and full line respectively after reference mapping, consensus mapping and Q-cpileup with QIT equals 26. After consensus mapping the coverage improved especially in less conserved regions of the virus. It is in these regions that the gain of iterative mapping is expected. (see peak around codon position 45). During mapping more mismatches are allowed compared to the default BWA setting. Currently the following parameters are used -n12 and -k6. Other settings might be more appropriate for other viruses. After applying Q-cpileup a dip in coverage is observed in the GC-rich region which is known to be error-prone and were a lot of false-positive findings are filtered out.

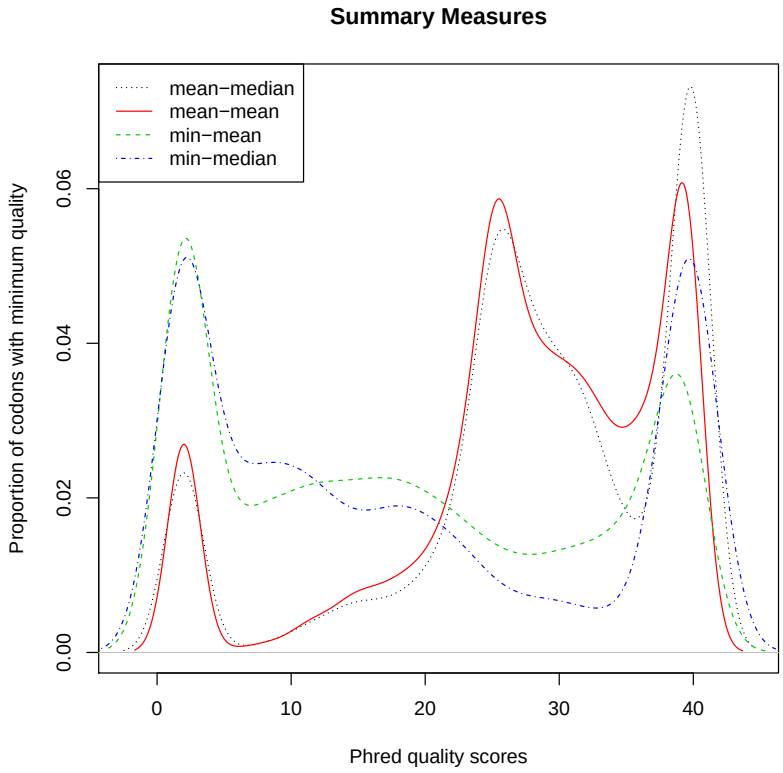


Figure A.2: Comparative analysis of different summarizations of the quality scores of codons. The quality scores need to be summarized at two stages. The quality scores of the three nucleotides building up a codon are summarized by their mean or by their weakest link, the minimum quality score of the three. The quality scores of identical codons at one position of the reference genome are summarized by their mean or by the median in order to get a quality measure in the final codon table. Each summarization is indicated with another color and line type. In the legend the summarization of the quality scores of the three nucleotides is stated first followed by the summarization of the quality scores of identical codons at one position. It is clear that the weakest link to summarize the quality scores of the three nucleotides gives the best separation between low and high quality codons. To summarize the codons at one position the mean is chosen as to take outliers into account.

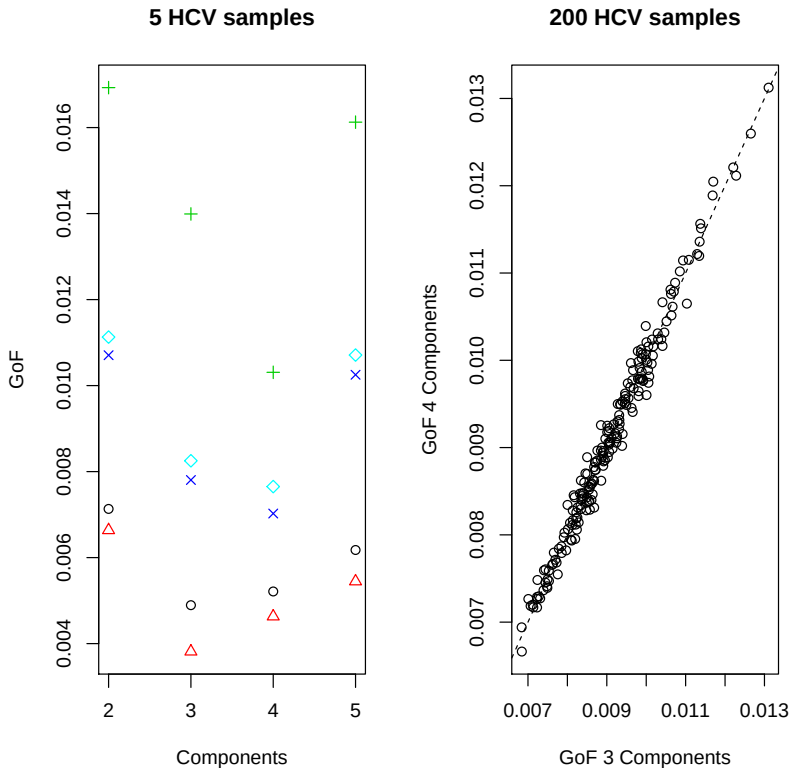


Figure A.3: a) Goodness of fit for mixture models with different number of components ranging from 2 to 5 calculated for 5 different HCV samples. Each sample is colored differently and has another plotting symbol. The fit with 3 or 4 components is the best depending on the sample. b) Goodness of fit for mixture models with 3 components on the x-axis and with 4 components on the y-axis for 200 HCV samples. The GoF is scattered around the identity line. No overall clear distinction between the two fits can be made and hence the one with the easiest biological interpretation is chosen, namely 3 components.

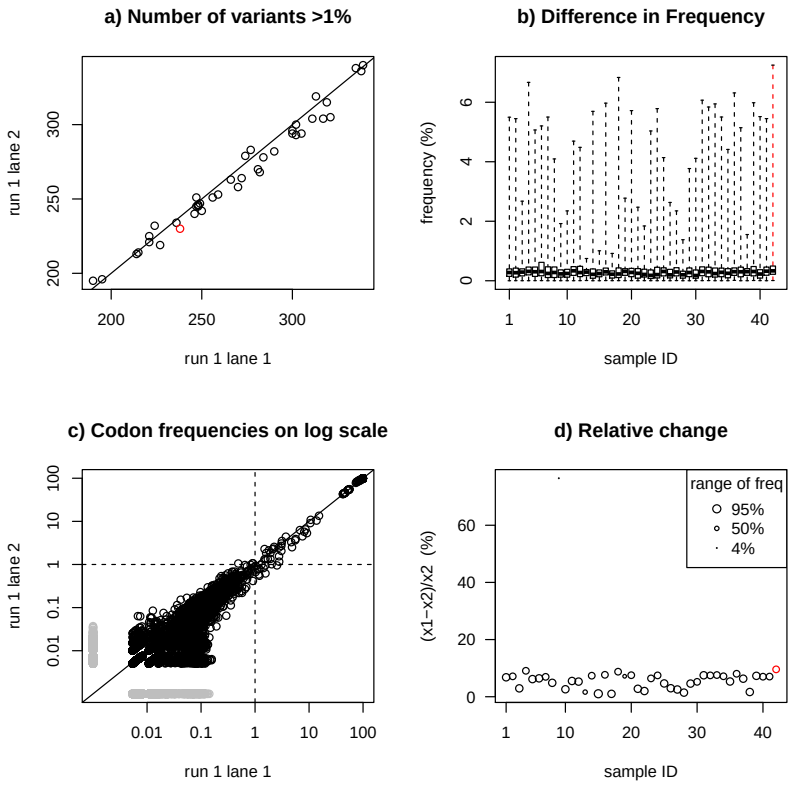


Figure A.4: Investigation of the effect of the intra-run variability of the QIT's on the reported codon frequencies when trimming but no filtering is applied. a) Plot of the number of variants with a frequency greater than 1% for each of the samples sequenced on lane 1(x-axis) and lane 2 (y-axis). b) Boxplots of the differences in codon frequency between the two lanes for each of the 42 samples. c) Comparison of all codon frequencies on the log scale between the two lanes for the sample where the maximum frequency difference is observed. The frequencies of codons not present in the other lane are plotted in gray. d) Relative change for these codons where the absolute difference was maximal is plotted for each sample and the sizes of the dots are scaled according to the estimated frequency in lane 1. The relative change is calculated $[x_1\text{-}x_2]/x_2$ with x_1 and x_2 the codon frequencies for lane 1 and 2 respectively. The sample where the maximum difference was observed is indicated in red.

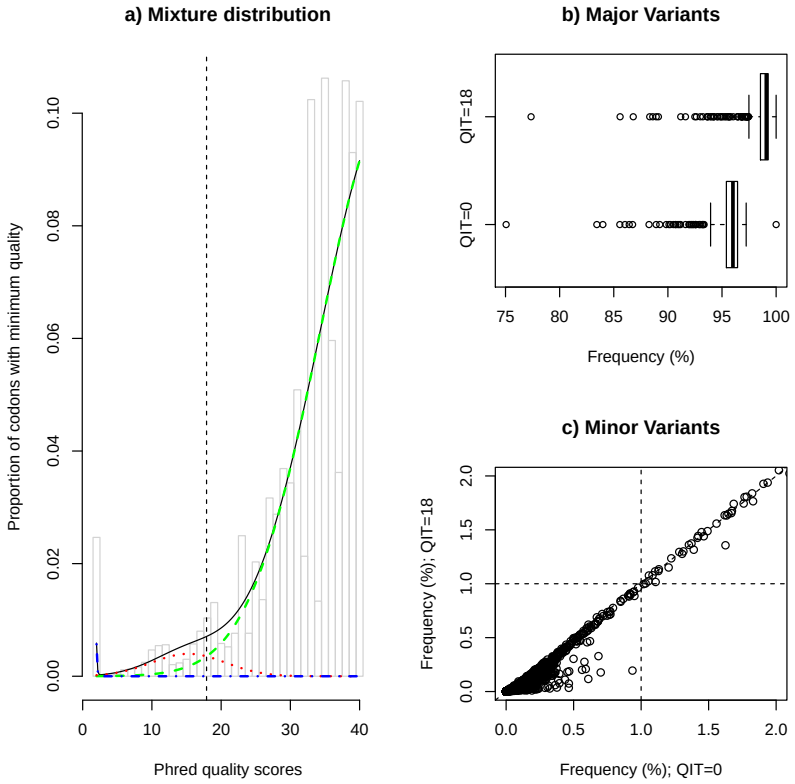


Figure A.5: Analysis of a HIV clinical sample a) Fit of mixture distribution with QIT indicated by verticaldashed line. b) Boxplot of the major variants frequencies before and after filtering based on QIT. By removing false positive variants, the major frequencies move towards 100% c) Scatterplot of minor variants with the frequencies before filtering on the X-axis and after Q-cpileup on the Y-axis. Again a reduction of false-positive findings after filtering is observed although less pronounced as compared to HCV samples.

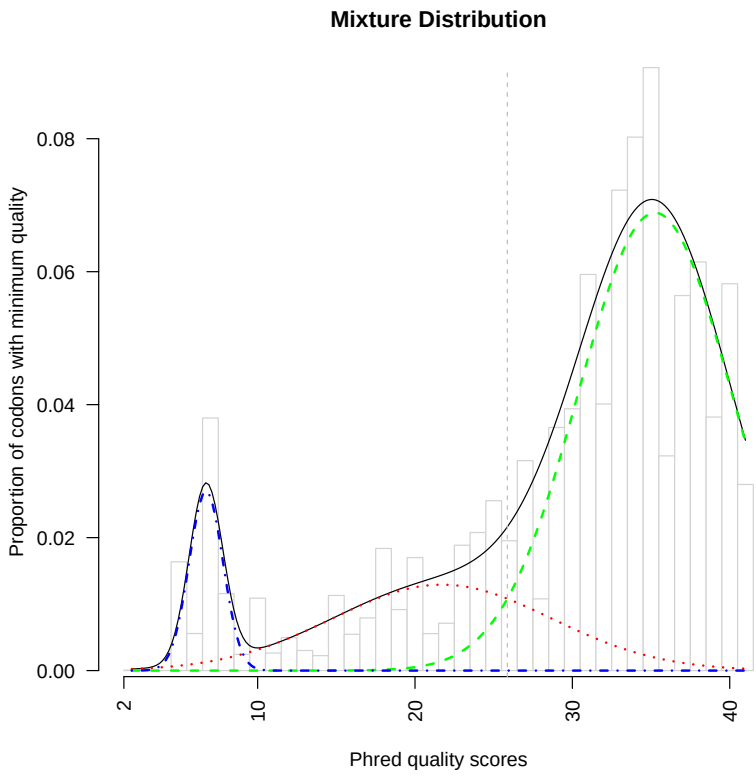


Figure A.6: Analysis of HCV sample sequenced on HISeq. Fit of mixture distribution with QIT indicated by dashed line which proves that the strategy as presented in chapter 3 can be used on HiSeq data as well. The quality scores in Illumina are calibrated using a quality table which is renewed each time a new chemistry is presented. Each time a few points of difference can be expected, but the automated pipeline should incorporate most of these. The points of differences observed are: The maximum quality score is 41 (while it was 40 before) resulting in 40 intervals of equal width. The point probability at 2 disappears as well, but a new mode can be seen around quality score 7. The automated pipeline checks if the first quantile of the total distribution is above 2. If not, a small initial estimate of the variance is chosen to force a point probability to be fitted, otherwise the same estimate for all three distributions is chosen.

B

Users Guide VirVarSeq

B.1 General descripton

VirVarSeq is a toolset designed to call variants at the codon level in viral populations from deep sequencing data. It is available, together with a users guide and test data at:

`http://sourceforge.net/projects/virtools/?source=directory.`

The pipeline starts from short-read sequences and a reference genome. It reports a codon table filtered based on the quality scores. A more detailed description along with the options will follow. The toolset consists of several components:

1. Mapping versus Reference (map_vs_ref.pl). Variant inference in viral populations starts by aligning the reads to a reference genome.
2. Determination of Consensus Sequence (consensus.pl). The reference genome may contain bases that do not represent the majority of the reads. Based on the previous alignment, the consensus at each position of the reference genome will be determined.
3. Mapping versus Consensus (map_vs_consensus.pl). The reads will be realigned against the consensus sequence to increase coverage.
4. Q-cpileup. A codon table will be constructed where the number of false-positive findings is reduced by exploiting the quality scores of the nucleotides. The method is adaptive to allow differences in qualities between the runs. It consists out of three consecutive analysis steps.
   (a) Retrieve the quality of codons (codon_table.pl)
   (b) Determine Q-intersection threshold (mixture_model.pl)
   (c) Filtering and reporting of codon table (mixture_model.pl)

B.2 Prerequisites

- BWA. BWA is a software package used for mapping of short-read sequences against a reference genome or the consensus sequence calculated in the pipeline. It is available at: http://bio-bwa.sourceforge.net/. VirVarSeq is recently tested with version 0.7.5a.
- R. R is a software environment for statistical computing used by Q-cpileup. It is available at: http://www.r-project.org/. The R-package rmgt is embedded in the pipeline. It is an R-wrapper to run the original Fortran code of McLachlan described in Biometrics [60] which fits truncated mixture models. The R-package is also used to produce some diagnostic plots to judge if chosen thresholds are acceptable and/or interpretable (more details in corresponding chapter 3). VirVarSeq is recently tested with version 3.0.1.
- Fortran. A fortran compiler is necessary to be able to run the R-package rmgt. A compiler can be downloaded from http://gcc.gnu.org/fortran/.
- Perl. Perl is a high-level, general-purpose, interpreted, dynamic programming language. It is available at: http://www.perl.org/. VirVarSeq is recently tested with version v5.10.1.
- Perl modules. The Perl module "Statistics::Basic" is used by VirVarSeq

B.3 Download

`https://sourceforge.net/projects/virtools/`

B.4 Getting started

1. Download VirVar.tar.gz. (see above)
2. Unzip VirVar.tar.gz.

   ```
   tar xvfz VirVarSeq.tar.gz
   ```

3. Install the R package "rmgt" (included in the distribution file VirVar.tar.gz). Within R use the following:

   ```
   install.packages("rmgt_0.9.001.tar.gz",repos=NULL,lib="lib")
   ```

4. Set environment variables so the VirVar Perl and R modules are found.

   ```
   export PERL5LIB=/home/ec2-user/VirVar/lib

   export R_LIBS_USER=/home/ec2-user/VirVar/R/lib
   ```

5. Test installation.

   ```
   ./run.sh
   ```

   The run.sh script creates a log file (VirVarSeq.log) where you can follow the progress.

6. Results. The output of the run.sh script is stored in directory <VirVar>/testdata/results The subdirectories are:

- <VirVar>/testdata/results/map_vs_ref
- <VirVar>/testdata/results/consensus
- <VirVar>/testdata/results/map_vs_consensus
- <VirVar>/testdata/results/codon_table
- <VirVar>/testdata/results/mixture_model

B.5 Usage

The VirVar pipeline makes use of a run.sh file.
User must specify:

- indir : directory where the fastq.gz files of the different samples to be processed are stored.
- outdir : directory where the output needs to be saved.
- samples : txt file with the names of the samples that need to be processed. The names are the first part of the fastq names created by Illumina. Multiple samples can be processed sequentially by having multiple sample names in the samples file.
- ref : path directing to the reference fasta file.
- startpos : position of the reference (at nucleotide level) where the determination of the consensus needs to start.
- endpos : position of the reference (at nucleotide level) where the consensus determination ends.
- region_start : position of the reference (at nucleotide level) where the pileup of the codons needs to start. This position is equal or higher to startpos.
- region_len : number of codon positions that need to be covered in the pileup table. Region_len 3 defines the length (at nucleotide level) on the reference genome that is reported in the pileup table.
- qv : a quality score used for filtering of the codon table. When applying Q-cpileup as described in chapter 3, Qv should be set to zero. The quality score used for filtering will be derived data driven. The option is left to specify a qv upfront. In this last case there is no need to run mixture_model.pl.
- trimming : If trimming is 0, soft-clipping as defined by the aligner will be ignored. If trimming is 1 (default), reads will be soft-clipped prior to the analysis.

B.6 FAQ

1. got message "Can‘t locate File.pm in @INC" running map_vs_ref.pl Please point the PERL5LIB environment variable to the <VirVar>/lib directory where the Perl libraries are installed
2. How can I install the Statistics::Basic Perl module? Perl modules can be installed using cpan. This installer will also install all dependend packages.

```
cpan -i Statistics::Basic
```

3. I don't have cpan, how can I install the Statistics::Basic Perl module? You must first install the cpan tool in order to install other Perl modules. Beware you must have root (or sudo rights) to install cpan tool.

```
sudo yum install cpan
```

B.7 Description test data

The test data provided together with the code, are 2 random samples containing 25% of the reads of a mixture of HCV plasmids (mixed 1:100). These mixtures of HCV plasmids are described in chapter 3 to assess the filtering accuracy of Q-cpileup, one of the components in the pipeline. Two different HCV plasmids were used for the mixture, each comprising the viral NS3-4A fragment. Site-directed mutations have been introduced into the con1b replicon plasmid pFK_i341_PI Luc_NS3-3'_ET (wild type) as described earlier [91, 92]. These plasmids, wild type and mutant, differ only in two codons (5 single nucleotides), as confirmed by Sanger sequencing. The HCV plasmid carrying the mutations is mixed into the wild type HCV plasmid at different proportions (1:10, 1:50, 1:100, 1:200). Following manufacturing protocols, the mixtures together with the WT and mutant are paired-end sequenced on Illumina GAIIx using 147 cycles. The fastq data from the 6 different samples can be downloaded from the European Nucleotide Archive, accession number PRJEB5028.

The reference genome used is hepatitis C virus type1b complete genome, isolate Con1 with GenBank ID AJ238799.1 which can be downloaded from:

`http://www.ncbi.nlm.nih.gov/nuccore/AJ238799`

B.8 Citing

Verbist B.M.P., Thys K., Reumers J., Wetzels Y., Van Der Borght K., Talloen W., Aerssens J., Clement L., Thas O. (2014) VirVarSeq: a low frequency Virus Variant detection pipeline for Illumina Sequencing using adaptive base-calling accuracy filtering. Bioinformatics, doi: 10.1093/bioinformatics/btu587.

C

Supplementary Information: Model Based Clustering

C.1 Sample preparation

The preparation of the HCV-NS3 plasmids as well as the HCV clinical samples can be found in the material and methods section of chapter 2. The sequencing protocols for Illumina and 454 are also described in this chapter 2.

C.2 Workflow

1. Off-line base calling with the option of second best base call:

   ```
   bustard.py --CIF <directory with intensities> --make --with
   -second-call --with-qseq --keep-dif-files
   ```

2. Demultiplex:

   ```
   configureBclToFastq.pl --input-dir <input directory>
   --output-dir <output directory>
   ```

3. Alignment with for instance BWA, creating a sam file
4. Convert second best base qseq files to fastq with fastqconverter from Casava:

   ```
   FastqConverter --in <inputfile> --out <outfile>
   ```

5. Add second best base call to sam, E2 and U2 tags are standard foreseen in sam to be filled by second best base calls and there quality scores: own Perl script submitted at sourceforge [http://sourceforge.net/p/vivambc/code/ci/master/tree/]
6. Convert sam to sorted bam with samtools:

```
samtools view -b -t <fasta of reference> <input sam file> |
samtools sort - <name of output>
```

7. Create heading with picard in order to be able to perform the next step:

```
java -jar /opt/picard-tools-1.86/AddOrReplaceReadGroups.jar
I=<inputfile>  O=<outputfile> LB=none PL=illumina PU=none
SM=none
```

8. Change bam positions using clipreads of GATK:

```
java -jar /opt/GenomeAnalysisTK-2.3-9/GenomeAnalysisTK.jar -T
ClipReads -I <input bam file> -o <output bam file > -R <input
ref file> -CR HARDCLIP\_BASES
```

9. Run R-script to perform model based clustering

C.3 R-code

The R-code together with the perl scripts are available under code at:

`http://sourceforge.net/p/vivambc/code/ci/master/tree/`

The R-code is parallelized which makes it possible to run each codon position on a separate core in order to speed up. The following command can be used:

```
mpirun -np <number of nodes> Rscript ViVaMBC.R.
```

The output is a codonTable.txt file where the position, the codons and their frequencies are reported. The code is tested on Amazon Web Services (AWS). For a region covering 181 windows the code runs for approximately 12 hours when a server with 16 cores and 60GB of RAM is used. Without parallelization it would take more than 7 days to obtain the results. Further optimalization of the R-code is most probably possible.

C.4 Error correction by second best base calling

The amount of errors that could be corrected by second best base calls is determined using the mixture of plasmids. These mixtures have two variant positions, position 36 (GTC (consensus) $\rightarrow$ ATG) and 155 (CGG (consensus) $\rightarrow$ AAA). All codons that are different from the two possible codons at each variant position are considered as error. For each false positive we check where the error occurs; which nucleotide or nucleotides within the codon differs from the ones in the true codons. In the next step, it is tested if the replacement of the error by the second best base call delivers one of the true codons. This procedure is repeated for the 4 mixing proportions, 1:200, 1:100, 1:50, and 1:10. In total 70% of the errors could be corrected by the second best base call (Figure C.1). The individual percentages for each codon position and each mixing proportion is shown in Table C.1.

C.5 Pileup

The results of ViVaMBC are compared with the codons present in the raw data after trimming (Table 2, chapter 4). The trimming is done by removing all bases, soft clipped by the alignment tool which is indicated in the CIGAR with S. These bases are most likely errors

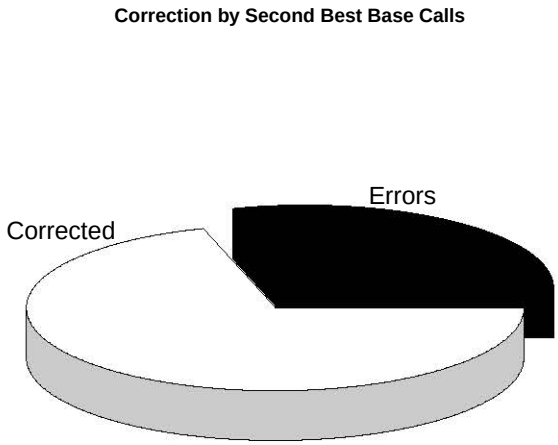


Figure C.1: Error correction by second best base calls. Pie charts where the total pie represents the number of errors observed. 70% of them could be corrected by the second best base call, while the other 30% remain error.

Pos	1:200	1:100	1:50	1:10
36	72.5	69.6	71.0	68.2
155	73.1	73.1	72.5	65.6

Table C.1: Percentage of errors that could be corrected by second best base calls for each variant position and for each mixing proportion.

since no translocations or big deletions are expected in the NS3 region. After trimming, all codons present in the data will be tabulated for each codon position. In chapter 4 this approach will be called pileup (in analogy with mpileup of samtools).

C.6 ViVaMBC at the SNP level

ViVaMBC has been optimized for $m = 3$ to retain linkage information between single nucleotide polymorphism, which allows for an immediate biological interpretation. Nonetheless, the algorithm can be applied with different window sizes m.

We have run ViVaMBC with window size $m = 1$ to call SNPs on the mixture of plasmids, in analogy with the existing methods. The estimated frequencies of the 5 known SNPs are reported in Table C.2. All variants could be retrieved with frequencies close to the mixing proportions. All other variants, besides the 5 SNPs, are assumed to be false-positive findings. ViVaMBC at the SNP level reports more false-positive findings compared with the existing methodologies similar to ViVaMBC at codon level (Table C.2). However, their frequencies remain well below 0.35%. Only one outlier was observed at a frequency of 0.65% for the mixing proportion 1:100. (Figure C.2). Overall ViVaMBC has a higher sensitivity and specificity for the discovery of SNPs down to a frequency of 0.5% in comparison with the other methods.

		ViVaMBC		
	SNP (WT)	1:200	1:100	1:50
36	A (G)	0.49	0.93	2.23
	T (T)			
	G (C)	0.43	0.88	2.18
155	A (C)	0.49	0.89	2.15
	A (G)	0.45	0.89	2.20
	A (G)	0.43	0.87	2.14
Number of false SNPs		132	139	209
Max Freq of false SNPs		0.32	0.65	0.34

Table C.2: Sensitivity and specificity of ViVaMBC at the SNP level. Frequency estimates of the true SNPs are close to the mixing proportions for all 3 mixes under investigation (1:200, 1:100 and 1:50). The bottom rows of the table report the total number of false SNPs over the whole NS3 region (543 bp long) together with their maximum frequency.

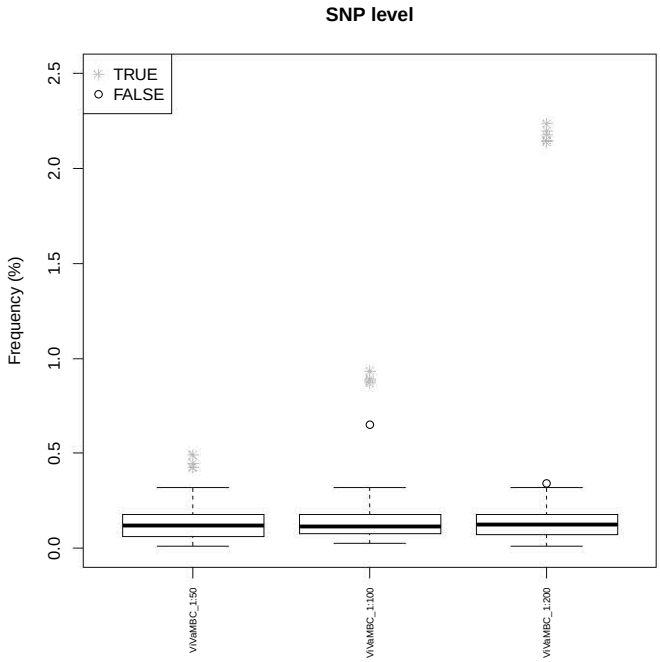


Figure C.2: ViVaMBC at SNP level Boxplots of the frequencies of all minor variants discovered in the three mixtures 1:200, 1:100 and 1:50 are plotted for ViVaMBC at SNP level. The true positives are indicated with gray crosses.

C.7 Contribution of second best base calls

ViVaMBC clusters variants based on the error probabilities of both the first and the second best base call. These second best base calls must be retrieved during base-calling. However, in some cases these second best base calls might be difficult to get, for instance sequencing providers perform often RTA and provide the best base calls only. Therefore, the performance of ViVaMBC is checked if only the best base calls are available. Equation 3 simplifies than to a simple logit model where c equals r instead of a multinomial logit model. Sensitivity and specificity is again investigated using the mixture of plasmids. The results are displayed in Table C.3. The estimated frequencies of the two real variants are close to the mixing proportions for the four different mixes. The number of reported variants over the whole NS3 region is displayed in the fourth column of Table C.3 together with the maximum frequency of the false-positive findings among the reported codons. For the plasmid mixtures, each codon differing from the wild type other than the spiked-in variant is considered to be an error. Hence, exclusion of second best base calls seems to have an increase in the specificity of the method without losing the sensitivity. However, some of the sequence differences at low-frequencies are expected to be real as they might originate from errors introduced during plasmid preparation. Hence, the increase in specificity is most probably a trade-off with the sensitivity for the very low-frequency variants.

Additionally, the influence of coverage depth on the accuracy of the frequency estimates is investigated using the plasmid data, mixed 1:200, at codon position 155. A similar experimental setup was performed as described chapter 4. The frequencies of the variants for this position for each of the 90 re-sampled datasets are plotted in Additional Figure C.3. The true codon variant AAA (green dots) was detected in all datasets. Averages frequency estimates over the 10 repeats are indicated with green triangles. The frequency estimate based on the error probabilities of best and second best base calls on the full dataset is indicated with a horizontal dashed line. In general, the frequency estimates are slightly underestimated. The number of false-positive findings is again much lower in comparison with ViVaMBC where the error probabilities of the second best base calls are taken into account. Although this might hint to an increased specificity, it is possible that it actually implies a decreased sensitivity as suggested above.

To investigate this further the method is applied on the GC-rich region of the clinical HCV sample used in chapter 4. The ViVaMBC results with and without second best base calls are plotted on the y- and x-axis respectively in Figure C.4. Codons that are exclusively reported with one of the methods are displayed in gray on the corresponding axis. The variants that were not present after 454 sequencing are displayed with triangles. Above 0.5% the two methods are in agreement, with slightly lower frequency estimates when omitting the second best base calls. Further, inclusion of the second best base calls in the model based clustering reveals more variants (similar to the results of the HCV plasmid) and sixteen of them are reported with the 454 experiment with frequencies up to 0.3%. Note, that the number of missed discoveries is probably higher since the 454 experiment was not sequenced deep enough to reveal frequencies below 0.05%. This suggests that some sensitivity is lost when second best base calls are omitted. However, both ViVaMBC implementations give very similar results when a reporting limit of 0.5% is applied.

C.8 Supplementary figures

Mixing Prop	36 ATG (%)	155 AAA (%)	$N°$ Codons	max noise freq (%)
1:200	0.44	0.40	289	0.52
1:100	0.89	0.84	291	0.51
1:50	2.20	2.16	301	0.57
1:10	10.82	9.89	253	0.38

Table C.3: Sensitivity and specificity of ViVaMBC in plasmid experiment when only the error probabilities of the best base calls are incorporated in the model. The estimated frequencies of the variants at codon position 36 and 155 for the four different mixing proportions are displayed together with the number of reported codons and the maximum frequency of the false-positive findings among them.

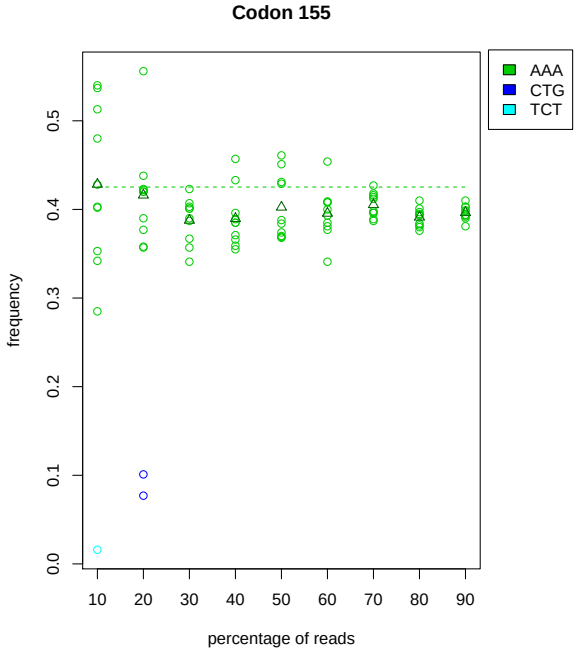


Figure C.3: Influence of coverage depth on the estimation of τ_j when ViVaMBC is solely based on the error probabilities of the best base calls. Datasets with lower coverages are generated by random sampling a fraction (f=0.1, 0.2, . . . , 0.8 ,0.9) of the reads from the original dataset. Ten datasets were generated for each fraction f resulting in 90 datasets. The reported variants for all re-sampled datasets were plotted and colored according to the discovered codon. The green dots indicate the true variant and the few others are false-positive findings. The average frequency of the true variant (averaged over the ten random samples) is indicated with triangles. The dotted line is the true frequency as estimated from the original dataset when the error probabilities of the second best base calls are taken into account.

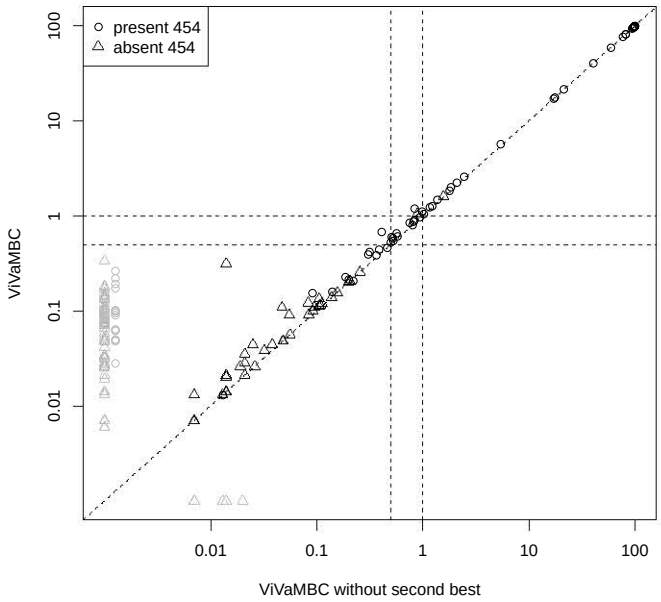


Figure C.4: Impact of the second best base call error probability on ViVaMBC. The frequency estimates of the codons revealed by ViVaMBC with and without second best base calls are plotted on the y-axis and x-axis respectively. Codons that are exclusively reported by one of the methods are plotted on the respective axis in gray. Codons represented with triangles were absent after 454 sequencing on the same sample and hence assumed to be false-positive findings. The reporting limits of 0.5% and 1% are displayed with dashed lines.

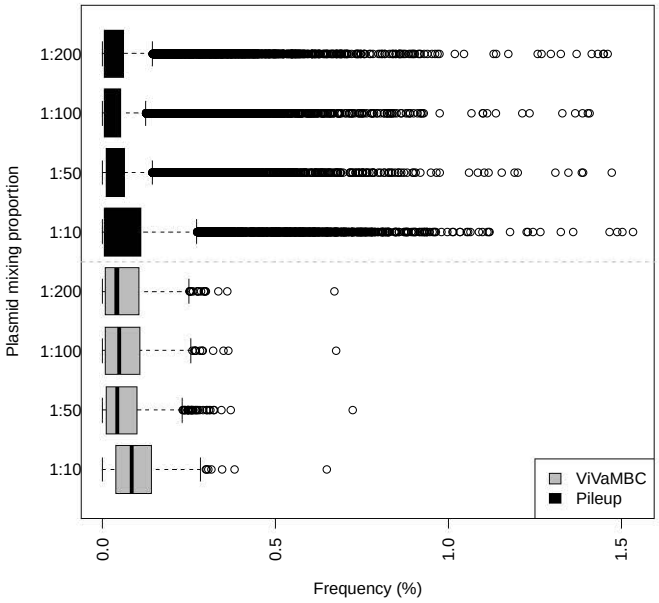


Figure C.5: Frequency distribution of the false-positive findings for the 4 mixing proportions after pileup and ViVaMBC. The raw data include some error induced variants with frequencies above 0.5% and even 1% which will hamper the discovery of the true low-frequency variants at position 36 and 155 (black boxplots). After applying ViVaMBC the frequencies of these false-positive findings only reach 0.25% (gray boxplots).

D

Curriculum Vitae

D.1 Experience

- Principal Biostatistician, Janssen Pharmaceuticals, april 2014 - Present.
  - Give statistical support to the discovery teams, with the main focus on oncology projects.
- PhD-student, Baekeland mandatory, Ghent University and Janssen Pharmaceuticals, 2011-2014.
  - Develop statistical algorithms for the detection of low-frequency variants within viral populations using Illumina sequencing technology which is presented in this thesis.
  - Data mining of three multi-dimensional data sources that are used throughout the drug discovery process to guide medicinal chemists: biological assay data, chemical structure information and transcriptomics. In this project, understanding the chemistry is key to be able to bridge the different data sources. (QSTAR: http://www.qstar-consortium.org)
- Scientific collaborator at Ghent University, 2010.
  - qPCR data analysis: Multiple imputation of missing data in housekeeping genes to allow a proper normalization of the data. Project in collaboration with University Hospital Gent.
- Post-doctoral scientist at Janssen Pharmaceuticals, 2005-2008.
  - Design, synthesis and validation of new biological active compounds within a therapeutic area in order to identify new entities for clinical evaluation.
  - Hit discovery by multivariate data analysis of biological assay data.

D.2 Education

- PhD Statistical Data Analysis (IWT Baekeland grant), Ghent University, 2011-present. Dissertation: Low-Frequency Variant Detection in Viral Populations using Massively Parallel Sequencing Data.
- Master of Statistical Data Analysis, Ghent university, 2008-2010. Dissertation: Chemosensitivity prediction by gene profiling: targeting biological relevance.
- PhD Medicinal Chemistry (IWT Scolarship), Catholic University Leuven, 2001-2005. Dissertation: Design and synthesis of potential β-turn mimetics: application in peptide chemistry and in the development of small molecules.
- Master of Chemistry, Catholic University Leuven, 1999-2001.
- Bachelor of Chemistry, Catholic University Leuven, 1997-1999.

D.3 Publications

- Thys K., Verhasselt P., Reumers J., Verbist B.M.P, Maes B., Aerssens J. Performance assessment of the Illumina massively parallel sequencing platform for deep sequencing analysis of viral minority variants. *under revision at Journal of Virological Methods*.
- Verbist B.M.P., Thys K., Reumers J., Talloen W., Aerssens J., Clement L., Thas O. VirVarSeq: a low frequency Virus Variant detection pipeline for Illumina Sequencing using adaptive base-calling accuracy filtering. 2014 *Bioinformatics*, doi:10.1093/bioinformatics/btu587.
- Verbist B.M.P., Clement L., Reumers J., Thys K., Vapirev A., Talloen W., Wetzels Y., Meys J., Aerssens J., Bijnens L., Thas O. ViVaMBC: estimating Viral sequence Variation in complex populations from Illumina deep-sequencing data using Model-Based Clustering. *under revision at BMC Bioinformatics*.
- Verbist B.M.P., Klambauer G., Vervoort L., Talloen W., QSTAR Consortium, Shkedy Z., Thas O., Bender A., Göhlmann H.W.H, Hochreiter S. Using Transcriptomics to Guide Lead Optimization in Drug Discovery Projects: Lessons Learned from the QSTAR Project. *accepted in drug discovery today*
- Mattiello F., Thas O., Verbist B. Principal Bicorrelation Analysis: Unraveling Associations Between Three Data Sources. *submitted to Journal of biopharmaceutical statistics*
- Perualila-Tan N., Kasim A., Talloen W., Verbist B., Göhlmann H.W.H., QSTAR Consortium, Shkedy Z. Joint modeling approach for uncovering associations between gene expression, bioactivity and chemical structure in early drug discovery to guide lead selection and genomic biomarker development. *submitted to Statistical applications in genetics and molecular biology.*
- Verbist B.M.P., Moerkerke B., Talloen W., Perera T., Göhlmann H.W.H., Goetghebeur E. Chemosensitivity Prediction by Gene Profiling: Targeting Biological Relevance. *submitted to journal of biometrics and biostatistics.*
- Verbist B. Minor Variant Detection In Virology with Model Based Clustering. Daghstühl Report, Computational Methods Aiding Early-Stage Drug Design (Dagstuhl Seminar 13212), 3:89. doi: 10.4230/DagRep.3.5.78.

- Gijsen H.J., DeCleyn M.A., Surkyn M., Van Lommen G.R., Verbist B.M.P., Nijsen M.L., Meert T., Wauve J.V., Aerssens J. 5-Sulfonyl-benzimidazoles as selective CB2-agonists – Part2. 2012 *Bioorg.Med.Chem.Lett*, 22:547-552.
- Gijsen H.J.M., Verbist B.M.P., Surkyn M. Fluoroalkyl subsituted benzimidazole cannabinoid agonists. 2009 *PCT Int. Appl.*, WO 2009077533.
- Bosmans J.-P. R.M.A., Berthelot D.J.-C., Pieters S.M.A., Verbist B.M.P., De Cleyn M.A.J. Equilibrative nucleoside transporter ENT1 inhibitors. 2009 *PCT Int.Appl.*, WO 2009062990.
- Gijsen H.J.M., De Cleyn M.A.J., Surkyn M., Verbist B.M.P. Benzimidazole cannabinoid agonists bearing a substituted heterocyclic group. 2008 *PCT Int. Appl.*, WO 2008003665.
- Verbist B.M.P., De Cleyn M.A.J., Surkyn M., Aerssens J., Nijsen.M., Gijsen, H.J.M. 5-Sulfonyl-benzimidazoles as selective CB2 agonists. 2008 *Bioorg. Med. Chem. Lett.*, 18:2574-2579.
- Kamoune L., De Borggraeve W.M., Verbist B.M.P., Vanden Broeck J., Coast G.M., Compernolle F., Hoornaert G. Structure based design of simplified analogues of insect kinins. 2005 *Tetrahedron*, 61:9555-9562.
- Verbist B.M.P., De Borggraeve W.M., Toppet S., Compernolle F., Hoornaert G.J. Development of new amino(oxo)piperidinecarboxylate scaffolds and their evaluation as β-turn mimics. 2005 *Eur.J.Org.Chem.*, 14:2941-2950.
- De Borggraeve W.M., Verbist B.M.P., Rombouts F.J.R., Pawar V.G., Smets W.J., Kamoune L., Alen J., Van der Eycken E.V., Compernolle F., Hoornaert, G.J. Design and synthesis of novel type VI-like β-turn mimetics. Diversity at the i+1 and the i+2 position. 2004 *Tetrahedron*, 60:11597-11612.
- Verbist B.M.P., Smets W.J., De Borggraeve W.M., Compernolle F., Hoornaert, G.J. Acid catalysed methanolysis of 2,5-diazabicyclo[2.2.2]octane-3,6-diones: scope and limitations 2004 *Tetrahedron Lett.*, 45:4371-4374.
- De Borggraeve W.M., Rombouts F.J.R., Verbist B.M.P., Van der Eycken E.V., Hoornaert G.J. Stereoselective intramolecular Diels-Alder reactions of 3-alkenyl(oxy)-2(1H)-pyrazinones. 2002 *Tetrahedron Lett.*, 43:447-449.